A Comparison of Data Envelopment Analysis and Stochastic Frontiers as Methods for Assessing the Efficiencies of Organisational Units

Laura Elizabeth Read

Presented for the qualification of
Doctor of Philosophy

Warwick Business School
University of Warwick

September, 1998

Supervisor
Emmanuel Thanassoulis

Synopsis

This thesis gives an overall view of the two most commonly used approaches for measuring the relative efficiencies of organisational units. The two approaches, data envelopment analysis (DEA) and stochastic frontiers (SF), are supposedly estimating the same underlying efficiency values but the natures of the two methods are very different. This can lead to different estimates for some, or all, of the units in an analysis.

By identifying the nature of these differences this work shows that it is possible to gain some insight into the nature of the underlying data and to say more confidently which of the two estimates is closer to the true efficiency for individual units.

In order to investigate the differences between the methods across different facets of the technology two important dimensions are chosen.

Firstly differences across scale size are investigated. It is shown how it is possible to define a measure of scale size in both the single output and multiple input and output cases. This measure of scale size can then be used to split the technology into regions of differing scale size enabling, for example, tests for the true nature of returns to scale in DEA. The measure of scale size developed in multiple dimensions necessitates a method for estimating an homothetic, constant returns to scale function.

Differences between the approaches across input mix are also investigated. These differences may highlight the abilities of the methods to correctly identify the elasticity of substitution between the inputs.

The results of the comparisons between the methods are summarised. This summary gives possible reasons for differences which may be found between the results of the two approaches, and an indication of what the nature of the estimates may be to the true efficiency values. An algorithm is then developed for using a comparison of the results from the two methods to help to identify the better estimates.

Contents

Tables

Chapter 3

Chapter 5

Chapter 7

Chapter 8

Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Chapter 8

Appendix 2

Acknowledgements

I would like to thank my supervisor, Emmanuel Thanassoulis, for all his support and encouragement during the process of my research and for giving me the chance to undertake this challenge. Without his thorough approach and our in-depth discussions, I would certainly not have the same confidence about my work.

Thanks to the DEA research group at Warwick for stimulating discussions and helpful comments: Rachel Allen, Claudia Sarrico, Ana Santos, Nikos Maniadakis, Ali Emrouznejad, Estelle Shale, Robert Dyson and Victor Podinovski. Also to Mark Freeman and Graham Sadler; thanks for helping with the conceptual and mathematical problems, but mainly for listening.

Most of all, I would like to thank my parents, for thinking this was a good idea, Bronya, and my grandparents. Especially Albert for being my inspiration.

Declaration

- This dissertation was written by Laura E. Read based on work undertaken by her at Warwick Business School.

- This work has not been accepted for any previous degree.

- This work has not yet been published.

Abbreviations and Notation

cmss	cross-mix scale size
COLS	Corrected Ordinary Least Squares
CRS	Constant Returns to Scale
DGP	Data Generating Process
DEA	Data Envelopment Analysis
DMU	Decision Making Unit
DRS	Decreasing Returns to Scale
E_{DEA}	efficiency of a unit estimated under DEA
E_{SF}	efficiency of a unit estimated under SF
E_{TRUE}	true efficiency of a unit
$\overline{E}_{DEA}$	average efficiency of all units estimated under DEA
$\overline{E}_{SF}$	average efficiency of all units estimated under SF
$\overline{E}_{TRUE}$	true average efficiency of all units
FD_j	Functional Deviation of DMU j
HSE	high scale efficiency
IRS	Increasing Returns to Scale
L(y)	Input set
LSE	low scale efficiency
MAD	Mean absolute deviation
MLE	Maximum Likelihood Estimation
mpss	most productive scale size
NDRS	Non Decreasing Returns to Scale
NIRS	Non Increasing Returns to Scale
OLS	Ordinary Least Squares
P(x)	Output set
PPS	Production Possibility Set
RTS	Returns to Scale

S(A)	cross-mix scale size of A
SF	Stochastic Frontiers
VRS	Variable Returns to Scale
x	input
y_{true}	true output - no inefficiency or random noise
$\tilde{y}$	efficient output - random noise, but no inefficiency
y_{obs}	observed output - random noise and inefficiency

Glossary

Constant Returns to Scale

A production frontier has CRS if, for an increase in all inputs by α%, all outputs increase by α%.

Decreasing Returns to Scale

A production frontier has DRS if, for an increase in all inputs by α%, all outputs increase by less than α%.

Increasing Returns to Scale

A production frontier has IRS if, for an increase in all inputs by α%, all outputs increase by more than α%.

Variable Returns to Scale

A production frontier has VRS if the frontier exhibits more than one of DRS, CRS or IRS.

Non Increasing Returns to Scale

If the production frontier only exhibits CRS and DRS, then it is said to be a NIRS frontier.

Non Decreasing Returns to Scale

If the production frontier only exhibits CRS and IRS, then it is said to be a NDRS frontier.

Functional Deviation

The functional deviation of DMU j (in an output orientation) is defined as the ratio of the true efficient output to the estimated efficient output.

Variation of Fit

Variation of fit is said to occur when there are regions of good and poor approximation to the true frontier.

Pure technical efficiency

The output technical efficiency of a DMU is a measure of how much the DMU could increase its outputs, keeping its inputs constant.

The input technical efficiency of a DMU is a measure of how much the DMU could decrease its inputs, keeping its outputs constant.

Scale efficiency

The scale efficiency of a DMU is a measure of how close the DMU is to operating at the mpss. It is a measure of how much more the DMU could increase its outputs than the level of pure technical efficiency, if it was operating at the most productive scale size.

Cross-mix efficiency

Output cross-mix efficiency is a measure of how much a DMU could increase its outputs, by changing its mix of inputs.

Full output technical efficiency

Full output technical efficiency is a measure of how much the DMU could increase its outputs if it was pure technically efficient, scale efficient and cross-mix efficient.

Allocative efficiency

Allocative efficiency gives a measure of the ability of the DMU to use inputs in the lowest cost mix, given the output level.

Inefficiency

Inefficiency is defined here as 1 - efficiency.

Elasticity of Substitution

The elasticity of substitution is defined as the ratio of the proportionate change in input proportions to the proportionate change in the slope of the isoquant. (The shape of the isoquant gives an indication of the elasticity of substitution. Very 'shallow' isoquants will have large substitution effects.)

Homothetic function

A homothetic Production function has isoquants that are radial projections of the unit isoquant.

True frontier

The true frontier is the frontier from which the units have been generated in the data generating process. Points on this frontier involve no inefficiency and no random noise.

Efficient frontier

The efficient frontier is the frontier that the units should be able to reach by eliminating their inefficiency. This frontier involves random noise.

Introduction

I. Introduction

In the late 1970's two classes of methods, **Data Envelopment Analysis** (**DEA**) (Charnes et al. (1978) and (1981)) and **Stochastic Frontiers** (**SF**) (Aigner et al. (1977), Meussen and van den Broeck (1977) and Battese and Corra (1977)) were developed for estimating the efficiency of organisational units (also called **decision making units** (**DMU**'s) or firms). These are units, such as schools or branches of a bank, which use the same set of inputs to produce the same set of outputs.

DEA is a non-parametric approach based on linear programming which takes the observed input and output values and forms a production possibility set[1] (PPS) making certain assumptions (see Banker, Charnes and Cooper (1984)). The distance of a DMU from the frontier of this set is then used as a measure of its inefficiency. This method gives an efficiency relative to the best practice DMUs. The SF approach, on the other hand, uses observed input-output correspondences to estimate an underlying relationship between the inputs and outputs. This function is then used as the frontier against which to measure the efficiencies.

[1] This concept will be defined in Chapter 1.

The methods have very different underlying structures which give rise to efficiency estimates which can differ between the methods. Currently, the choice of which method to use is often dependent upon which one is seen as the easiest to implement rather than any reasoned argument for the better performance of the chosen method. This leads to DEA often being chosen in preference to SF methods (although there are other reasons for preferring DEA including the fact that the results can be easier to analyse). The estimates given by the SF method are conditional on the total error[2] and this can be used as a reason not to use the SF method - Banker et al. (1988); "...[with SF estimation] we encounter problems with lengthy algorithms for estimation and difficulty in isolating estimates for individual observations." However, the software now available makes it possible for the SF estimates to be obtained relatively easily and as we will see, the estimates are very good when the assumptions of the methods are met.

Unfortunately there is no easy answer as to which of the two approaches performs better: The performance of the methods is highly dependent upon the data set which is being analysed. In some data sets one of the methods will give better estimates for all the units; and in others, some of the units will be given better estimates under one method and others, better estimates under the second method. If both methods are applied to the same data set, there must be some way to

[2] See Chapter 1, Section 1.3.

explain the differences and similarities between the estimates in order to validate the results. It is proposed here that a comparison between the results of the methods can be used to obtain a view as to which of the methods is more likely to be giving the better estimates both across the whole technology[3] and for specific regions of the technology.

II. Preliminary research objective

The objective of this research is to investigate the sources of the similarities and differences between SF and DEA based estimates of efficiency so that they can be exploited in applications of the methods. In order to be able to make a judgement about which of the methods to use on a particular data set, the performance of the methods will be analysed for several simulated data sets. Two questions will be addressed: Firstly, is there any way to look at the comparative performance of the methods to infer any properties of the data set which affect the performance of the method; and secondly, if we can use the methods to identify properties of the data set, can we proceed to state which of the methods is likely to be outperforming the other?

The main focus of the comparison will be between DEA and SF methods in assessing firm specific technical efficiency. Forsund (1992) questions whether it is reasonable to compare a deterministic method with a stochastic method. However, this comparison is chosen here,

[3] The technology will be defined in Chapter 1.

not only because these two methods are the most widely used, but also because the differences in their assumptions can give indications as to which method is outperforming the other when the estimated efficiency values differ. The results presented here are aimed at obtaining a more informed judgement as to the suitability of the methods for analysing specific data sets.

III. Secondary research objective

The differences between the methods will vary across the technology. How this variation occurs will depend upon which of the underlying assumptions of a method is not met by the data. It is proposed here that there are two main dimensions in which this variation can occur: variation across the input or output mix (i.e. in the two-input case, the variation as one input increases and the other decreases) or variation across scale size.

In order to measure how the differences between the estimates vary across scale size it is necessary to define a measure of relative scale size and this is done in the central chapters of the thesis (Chapters 4 and 6).

It is shown in Chapter 4 that the Malmquist input quantity index can be used to measure scale size, first in the single-input, single-output case and then in a multiple-input, single-output case. The distance functions

in the index must be measured against a CRS frontier in order for the relative scale sizes to be independent of the reference output level.

In Chapter 6 we find that this method does not easily generalise to the multiple-output case. In order to measure relative scale sizes in the case of multiple inputs *and* outputs, the Malmquist input quantity index can be used only if the distance functions in the index are measured against an homothetic CRS frontier.

IV. The structure of the thesis

The next chapter will give an introduction to the methods used to measure technical efficiency and their comparative advantages and disadvantages. The second chapter outlines the reasons for differences between the estimates from the methods and gives several hypotheses for the effect on the estimates of certain types of underlying technologies. These hypotheses are then tested in the following five chapters. Chapter 8 summarises the conclusions about how the methods perform and develops an algorithm to show the information that can be gained by comparing the results from the two methods. Finally Chapter 9 presents the conclusions. Technical details of the Stochastic Frontier method can be found in Appendix 1. Appendix 2 outlines the data generating process for four different sets of data, which will be used throughout the thesis. Appendix 3 gives the

hypothesis tests which are used in Chapter 5 and Appendix 4 expands on some of the technical details used in Chapter 6.

Chapter 1

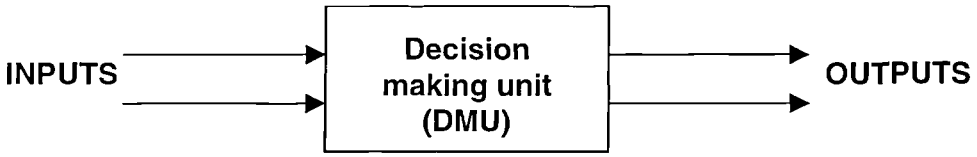
The measurement of efficiency

1.1 Production theory

1.1.1 Production processes

Production is any process that converts a set of inputs into a set of outputs.

Figure 1-1. The production process



For example, a factory uses inputs of raw materials, labour, operating costs and produces goods; a school uses inputs of pupils with certain obtained levels of achievement, operating expenses and teachers of varying qualifications and skills, and would like to maximise the academic and other attainments of the pupils. Any process taking a set of inputs to produce certain outputs can be viewed in this way. It may be difficult to define the inputs and outputs or to measure them, but once these difficulties have been overcome, it is not necessary to know about the actual processes involved in converting the inputs into the outputs in order to measure how well the units are performing: Instead a set of similar units can be taken and compared. In a chemical process the inputs are converted into outputs in a predictable way - the relationship between the inputs and outputs has a precise functional form. However, in other production processes the conversion of inputs

into output does not generally follow a known functional form (e.g. the school example above). This means that it is not possible to know exactly what the maximum output obtainable from the given inputs is. The maximum output has to be estimated from the observed data. This is the difference between an engineering definition of efficiency and the relative efficiency estimated in production theory. Production is now used to mean any conversion process where the outputs are to be maximised subject to a fixed set of inputs[1].

1.1.2 Inputs and outputs

In some cases the factors involved in a production process are obvious, for example, when building a house the inputs would be the raw materials used and the labour. However, in many cases the choice of inputs and outputs is not obvious, e.g. assessing the efficiency of banks, schools, countries, etc. When inputs and outputs are not easily measurable, proxy variables may be chosen to represent them. For example, Thanassoulis and Dunston (1994) use the percentage of pupils not taking free school meals as a proxy for the socio-economic background of the pupils in a school.

[1]This is the output-oriented view of production. Production can equally well be defined as a conversion process where the inputs are to be minimised subject to a fixed set of outputs. The output orientation will be used in this thesis.

In all cases it is important to be clear about the process which is being investigated. For example, when measuring the efficiency of a school, is the sole aim of the school to maximise academic attainment or should the non-academic attainment of the pupils be considered as well?

There may be other problems defining inputs and outputs, such as variables which are anti-isotonic, that is, inputs (outputs) which it would be preferable to increase (decrease), e.g. pollution is an anti-isotonic output (see Athanassopoulos and Thanassoulis (1995) for another example). Another problem that may occur is the presence of categorical variables, i.e. variables that can only take certain values (Banker and Morey (1986)).

Once the inputs and outputs have been decided upon, in order to be able to measure efficiency, a benchmark is needed. This is given by the production frontier.

1.1.3 The production frontier

Two approaches are used to define the production frontier in the literature. The first, the Neo-classical Approach (Frisch (1965)) specifies the production or transformation function and dual cost function immediately. The **production function** is a mathematical

representation of the transformation between inputs and outputs and is defined as the maximum possible output obtainable from given inputs. It follows that observations may only lie below a production function.

The second approach, the Axiomatic Approach (Koopmans (1957), Debreu (1959), Shephard (1970)), is based on production sets. This is a broader approach and easily incorporates multiple inputs and outputs. These two approaches are equivalent.

The neo-classical approach is parametric - a specific form must be given for the production function. A non-parametric frontier has no assumption of functional form and the frontier is formed using the axiomatic approach.

The next section gives some of the important definitions of the axiomatic approach which will be referred to throughout the thesis.

1.1.3.1 **The axiomatic approach** (Shephard (1970))

Consider a production process in which m inputs are converted into s outputs. Let $\mathbf{y} \in \Re^s$ denote the vector of outputs and $\mathbf{x} \in \Re^m$ the vector of inputs. The **technology**, **graph** or **production possibility set** (PPS) is given by

$$PPS = \{(x,y) \in \Re_+^{m+s} : x \text{ can produce } y\}. \qquad (1\text{-}1)$$

This is the set of all possible input-output combinations and is formed using certain axioms, e.g. Shephard (1970) or Färe and Primont (1995). The technology can equally well be described by the input or output sets. An **input set**, L(**y**), of a technology is the set of all input vectors **x** yielding at most, output **y**.

$$L(y) = \{ x \in \Re_+^m : y \text{ can be produced by } x \} \qquad (1\text{-}2)$$

Similarly, the **output set**, P(**x**), is the set of all output vectors **y** which can be produced by the input vector **x**.

$$P(x) = \{ y \in \Re_+^s : x \text{ can produce } y \} \qquad (1\text{-}3)$$

The **isoquant** corresponding to an output **y** > **0** is a subset of the boundary of the input set L(**y**) defined by

$$\text{Isoquant} = \{x : x \geq 0,\ x \in L(y),\ \lambda x \notin L(y) \text{ for } \lambda \in [0,1)\} \qquad (1\text{-}4)$$

The efficient subset[2] $E(\mathbf{y})$ of an input set $L(\mathbf{y})$ is given by

$$E(y) = \{x : x \in L(y), x' \leq x, x' \neq x \Rightarrow x' \notin L(y)\}. \qquad (1\text{-}5)$$

For the single-output case[3] the production function is then the maximum output attainable from given inputs

$$f(x) = \max \{ y \in \Re_+ : y \in P(x) \}. \qquad (1\text{-}6)$$

This function may be specified parametrically or it may be formed from the observed input-output correspondences.

1.1.3.2 Parametric and non-parametric frontiers

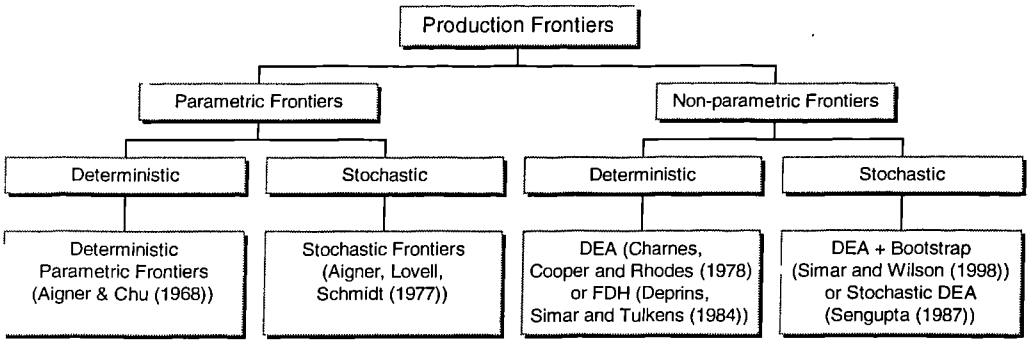
A parametric frontier has a precise mathematical form. A non-parametric frontier is formed using certain assumptions about the nature of the technology. Parametric and non-parametric frontiers can be subdivided into frontiers that are stochastic or deterministic (see Figure 1-2).

---

[2] Note that this is not the same as the isoquant. The isoquant may contain sections that are parallel to the axes. These cannot be part of the efficient frontier.

[3] For a generalisation to multiple outputs see Chapter 6.

Figure 1-2. Production frontiers



The stochastic case assumes that it is not possible to fully specify the function and allows for random noise. The deterministic case assumes away any random factors.

The most common methods for efficiency estimation are DEA in the non-parametric literature and SF in the parametric literature. These will be outlined in Sections 1.2 and 1.3 of this chapter and the rest of the thesis will concentrate on them. Both DEA and SF approaches are derived from the methods of measuring efficiency introduced by Farrell in 1957 who suggested measuring the efficiency of a firm relative to an empirical production frontier.

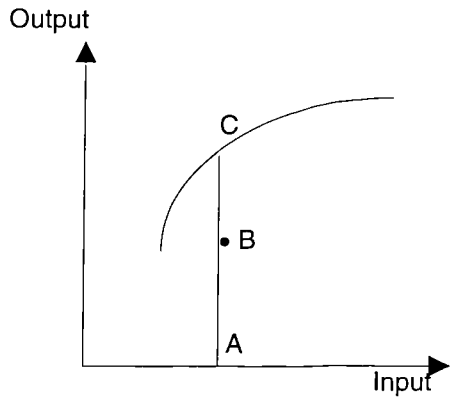
Before considering the methods in detail, the different ways that efficiency can be defined will be examined.

1.1.4 Technical efficiency

Once a production frontier has been estimated, the deviation of an individual firm's output from the maximum output which it could have achieved given its inputs (the corresponding point on the production frontier) can be used to define a measure of the technical efficiency[4] of the firm. The amount by which the observation lies below the production frontier can be regarded as a measure of its inefficiency.

The **technical efficiency** of a firm is defined here to be the ratio of the observed output to the efficient output. This measure necessarily has

Figure 1-3. Measuring technical efficiency.



[4] This is the output technical efficiency. It will be defined here for the single-output case and generalised later to the multiple-output case.

values between zero and one. If a firm has a technical efficiency of 0.7, it means that the firm is producing 70% of the output that it could produce if it were fully efficient.

For example, in Figure 1-3, the output efficiency of DMU B is given by

$$\text{Output technical efficiency of DMU B} = \frac{\text{observed output}}{\text{efficient output}} = \frac{AB}{AC}. \quad (1\text{-}7)$$

The **inefficiency** of the firm can be, and is, defined in several different ways. For example, the inefficiency could be defined as the inverse of the efficiency, or more commonly, the inverse of the efficiency minus one (Banker et al. (1988)). In the previous example, this definition of the inefficiency will give

$$\text{Inefficiency of DMU B} = \frac{AC}{AB} - 1 = \frac{AC - AB}{AB} = \frac{BC}{AB}. \quad (1\text{-}8)$$

This is the percentage by which the observed output would need to increase for the DMU to become efficient.

Alternatively, the definition of inefficiency that will be used in this thesis is one minus the efficiency.

$$\text{Inefficiency of DMU B} = 1 - \frac{AB}{AC} = \frac{AC - AB}{AC} = \frac{BC}{AC}. \qquad (1\text{-}9)$$

This is the percentage of the efficient output level that the unit is wasting due to inefficiency.

1.1.5 Scale efficiency

So far we have only been considering technical efficiency, that is, a measure of how far an observation is away from the production frontier. However, once a unit has reached the production frontier, it still may not be efficient - all points on the production frontier are not equally productive unless the whole frontier has constant returns to scale.

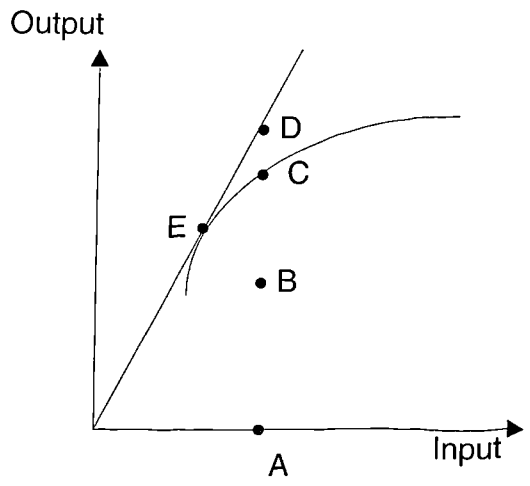
The **productivity** of a unit can be defined as the amount of output that a unit produces given a unit of input. In the single-input, single-output case this is defined as

$$\text{Productivity} = \frac{y}{x} \qquad (1\text{-}10)$$

Now, the productivity can change across the frontier, but there will always be at least one point for each input-output mix that is operating at the greatest productivity level[5]. This is known as the **most productive scale size** (**mpss**, Banker (1984)). This is the part of the frontier for which the tangent hyperplane through the origin has the greatest gradient.

In Figure 1-4, the mpss is at point E. All points on the frontier, which are not operating at the mpss, are inefficient. This inefficiency is known as **scale inefficiency**.

Figure 1-4. Most productive scale size



[5] Assuming that the PPS is closed and bounded.

The scale inefficiency of a point on the frontier[6] is a measure of how far away the frontier at that point is from the CRS frontier. In Figure 1-4, DMU B is technically inefficient as it is operating below the efficient frontier. If it increases its output level to become technically efficient it will reach point C. However, at C it would not be operating at the mpss: It would be scale inefficient. The scale efficiency of C is defined as

$$\text{scale efficiency of point C} = \frac{AC}{AD}. \qquad (1\text{-}11)$$

The **full technical efficiency** is a measure which incorporates both technical and scale efficiencies. This is generally[7] defined as

Full technical efficiency of DMU B

$$= \text{technical efficiency x scale efficiency} \qquad (1\text{-}12).$$

$$= \frac{AB}{AC}\frac{AC}{AD} = \frac{AB}{AD}$$

[6] Note that scale inefficiency is only defined for points on the frontier.

[7] In Chapter 6, it will be shown that this measure should also incorporate cross-mix inefficiencies.

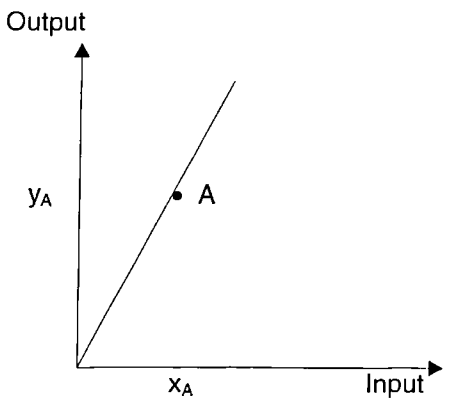
Other types of efficiency can be considered, such as allocative efficiency, but they will not be dealt with here.

1.1.6 Returns to Scale

Related to scale efficiency is the concept of Returns to Scale.

The **Returns to Scale** (RTS) of a point on the production frontier are defined as the amount that all the outputs will increase by for a proportionate increase in all inputs. That is, if all inputs increase by 1%, how much can all the outputs increase by? Note that this is **not** given by the gradient of the production frontier at that point. The outputs will all increase by 1% if the tangent hyperplane to the frontier at that point goes through the origin. (See Figure 1-5 for the single-input, single-output case.) This is called **constant returns to scale**.

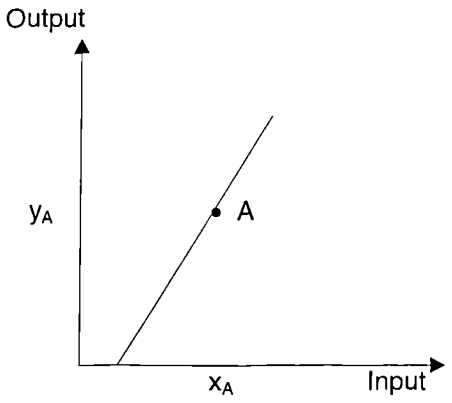
Figure 1-5. Constant returns to scale



In Figure 1-5, DMU A uses input x_A to produce output y_A. If the inputs of A are increased by 10%, then the outputs must increase by 10% in order for the unit to remain efficient.

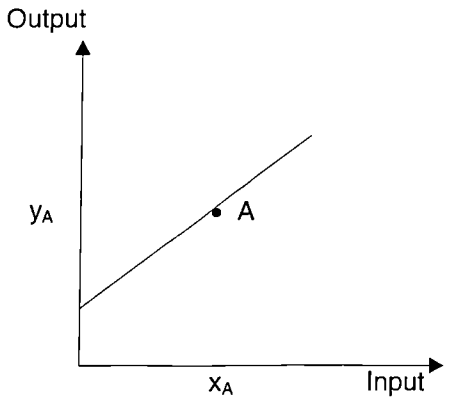
If an increase in all the inputs by 1% leads to an increase in all the outputs by more than 1%, we say that the frontier at this point is exhibiting **increasing returns to scale**. This is equivalent to the tangent hyperplane at the frontier point having a negative intersection on the output axis. The single-input, single-output case is illustrated in Figure 1-6.

Figure 1-6. Increasing returns to scale



Conversely, for a tangent hyperplane with positive intersection on the output axis, we have **decreasing returns to scale**, see Figure 1-7.

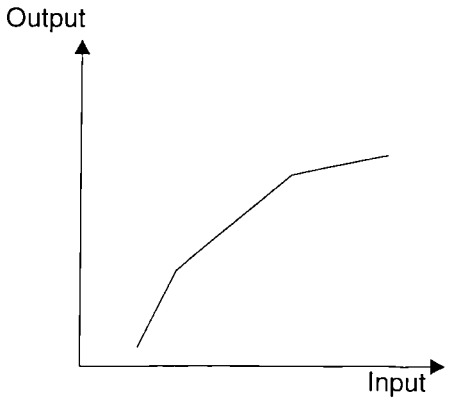
Figure 1-7. Decreasing returns to scale



This is equivalent to saying that, for an α% increase in all the inputs, the outputs increase by less than α%.

These concepts have been illustrated in Figures 1-5 to 1-7 for frontiers that are globally CRS, IRS or DRS. However, it is possible for the frontier to exhibit these characteristics locally, see Figure 1-8.

Figure 1-8. Variable returns to scale



In this example, the frontier exhibits first IRS, then CRS and finally DRS as the input increases. We will use the term **variable returns to scale** to denote any frontier for which CRS does not hold.

Scale inefficiency will become important in Chapter 5 of the thesis where differences between DEA and SF approaches are investigated across scale size. The following sections describe the two methods for efficiency measurement which will be investigated.

1.2 Non-parametric frontiers

Rather than explicitly stating the form of the frontier, non-parametric methods estimate the frontier using the data. The data is used to form a production possibility set and the frontier of this set is used as the benchmark.

The non-parametric model that we consider, is Data Envelopment Analysis (DEA) which was developed by Charnes, Cooper and Rhodes (1978).

1.2.1 Data Envelopment Analysis

The DEA method is based on the Axiomatic Approach outlined earlier. The PPS is formed by making certain assumptions (see Charnes,

Cooper and Rhodes (1978), Banker, Charnes and Cooper (1984) and Olesen (1995)).

Consider a set of n DMUs with m inputs and s outputs. Let y_{rj} denote the level of output r and x_{ij} the level of input i for DMU j.

The output efficiency of DMU 0 (where DMU 0 denotes the unit in the collection of DMUs j = 1,...,n which is being assessed) is $1/\theta_0^*$ where θ_0^* is the optimal value of θ_0 in the following model:

Model 1. Data Envelopment Analysis

$$\max \left(\theta_0 + \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right)\right)$$

subject to

$$\sum_{j=1}^{n} x_{ij}\lambda_j = x_{i0} - s_i^- \qquad (i = 1,2,...,m)$$

$$\sum_{j=1}^{n} y_{rj}\lambda_j = \theta_0 y_{r0} + s_r^+ \qquad (r = 1,2,...,s)$$

$$\lambda_j,\ s_i^-,\ s_r^+ \geq 0\ ,\ \forall\ j, i, r,\ \theta_0 \text{ unconstrained,}$$

where y_{rj} is the level of output r and x_{ij} the level of input i for DMU j, and ε is a vanishingly small positive number.

The extra term $\varepsilon(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+)$ ensures that any facets which are parallel to the axes are not given efficiency values of 1 (see footnote 2).

Model 1, due to Charnes, Cooper and Rhodes (1978), assumes that efficient production is characterised by constant returns to scale. The optimal value of θ_0 measures the ratio of the efficient output to the observed output.

Banker, Charnes and Cooper (1984) modified model 1 to include variable returns to scale by adding the constraint

$$\sum_{j=1}^{n} \lambda_j = 1. \qquad (1\text{-}13)$$

This constraint removes the CRS assumption by restricting the PPS to the convex hull of the observed DMUs. This model, known as the BCC or VRS model, allows for local increasing, constant and decreasing returns to scale.

Similarly models capturing non-increasing returns to scale[8] (NIRS) and non-decreasing returns to scale[9] (NDRS) can be created by replacing (1-14) by

$$\sum_{j=1}^{n} \lambda_j \leq 1 \qquad (1\text{-}14)$$

for NDRS and

$$\sum_{j=1}^{n} \lambda_j \geq 1 \qquad (1\text{-}15)$$

for NIRS.

Note that it is not possible to specify a DRS or an IRS DEA model as there will always be at least one point on the DEA frontier which has constant returns to scale, i.e. the mpss.

Various extensions have been made to these basic models. See Charnes et al. (1995) and Cooper, Thompson and Thrall (1996) for

[8] i.e. a frontier which can exhibit IRS and CRS but not DRS.
[9] i.e. a frontier which can exhibit CRS and DRS but not IRS.

discussions of these extensions. Also, Seiford (1996) gives an overview of the important developments by considering the 'state of the art' in four different years; 1980, 1985, 1990 and 1995.

One important development was the Free Disposal Hull method (Deprins, Simar and Tulkens (1984) and Tulkens (1993)) which allows for the relaxation of the convexity assumption in DEA. This assumption will be considered in detail in Chapter 2.

1.3 Parametric frontiers

A parametric frontier model depends on specifying a functional form which relates the outputs to the inputs and then estimating the parameters of this function using one of the methods outlined in Appendix 1 subject to certain assumptions about the distribution of the residuals.

In 'normal' parametric production frontier models it is only possible to consider a single output with multiple inputs or a single input with multiple outputs. However, there are now several methods for incorporating multiple variables into a SF approach. These will be discussed in detail in Chapter 6.

1.3.1 Models which do not allow for random noise

Farrell (1957) suggested specifying a particular functional form for the frontier but it was not until Aigner and Chu (1968) that this idea was explored. Until then, all econometric estimation of production relationships had involved estimating an 'average' production function - allowing observations to lie above and below the estimated function. Aigner and Chu (1968), Afriat (1972) and Richmond (1974) all assume a production function of the form

$$y_{eff} = f(\mathbf{x};\beta) \qquad (1\text{-}16)$$

where y is the maximum possible output for the given input vector **x**. Schmidt (1976) explicitly specified the frontier production function as:

Model 2. The Deterministic Production Frontier

$$y_{obs} = f(\mathbf{x};\beta) - u, \qquad u \geq 0$$

where y is the output, **x** is the vector of inputs and u is the residual. β is a vector of parameters estimated by the method.

In this model the whole of the residual is counted as inefficiency. The parameters β can be estimated by linear or quadratic programming (Aigner and Chu (1968)), corrected ordinary least squares (COLS) (Olson, Schmidt and Waldman (1980)) or maximum likelihood (Afriat (1972), Schmidt (1976)), and the function is chosen to have as much flexibility as possible. The technical efficiency of each observation can then be computed directly from the residual.

For details of how to estimate parametric frontiers see Appendix 1.

There are some problems with deterministic frontiers; firstly, the estimates of the parameters have no statistical properties as assuming a one-sided disturbance violates the regularity conditions for maximum likelihood estimation (see Schmidt (1985)). Secondly, the residuals, i.e. the differences between the estimated efficient outputs and the observed outputs, are taken as measures of efficiency. This means that all variation in firm performance is attributed to variation in firm efficiencies relative to the production frontier. However, there are several reasons other than technical inefficiency for firms to lie below the production frontier;

1. "...weather, unpredictable variations in machine or labor performance, and so on." Zellner, Kmenta and Dreze (1966).

2. "Pure random shocks in the production process. For example some parts of a product may be damaged through careless handling; or some products are defective, etc." Aigner and Chu (1968).
3. "...differences in economic efficiency. Given a production function and the market situation, the firm should produce a certain level of output so as to maximize its profits." Aigner and Chu (1968).
4. "...luck, climate, topography, and machine performance. Errors of observation and measurement on y constitute another source..." Aigner et al. (1977).
5. "definitional and measurement problems in the variables" Timmer (1971).
6. Functional misspecification (i.e. using a functional form in the method which is too restrictive to capture all the properties of the true relationship between inputs and outputs).
7. Incorrect assumptions about the distribution of the inefficiency. This will be illustrated in Chapter 3.

If one of these reasons, i.e. functional misspecification, only occurs in the parametric frontier method, then a comparison with the DEA results may highlight this as the reason for the difference between the results. It may then be possible to decide which of the methods is giving the better estimates. This is the focus of the thesis and will be examined in more detail in the next chapter.

1.3.2 Models which do allow for random noise

The Stochastic Frontier Model

The stochastic frontier model was first proposed by Aigner, Lovell and Schmidt (1977), Meeusen and van den Broeck (1977) and Battese and Corra (1977). It was developed to introduce random factors by fitting a production function, and allowing the frontier to shift around the fitted function for individual companies. This is done by using a composed error term. The error term is split into a one-sided error term measuring firm-specific inefficiency and a two-sided error term showing random fluctuations, which is identically and independently distributed across firms.

Consider the model:

Model 3. The Stochastic Production Frontier Model

$$y_i = f(\mathbf{x}_i ;\beta) + \varepsilon_i \qquad \text{where } \varepsilon_i = v_i - u_i, \quad i = 1,...,n$$

v_i represents the symmetric disturbance (all events which are not under the control of the firm). $\{v_i\}$ are assumed to be independently and identically distributed as $N(0,\sigma_v^2)$.

u_i represents the inefficiency (all the events which are under the control of the firm). u_i is assumed to be distributed independently of v_i and $u_i \geq 0$.

The efficient production frontier is now $f(\mathbf{x}_i; \beta) + v_i$

The stochastic production function can be estimated by COLS or maximum likelihood (for details see Appendix 1). The distribution of u must be specified in both cases.

Consider for simplicity a log-linear model (the Cobb-Douglas function)

$$\ln y_i = \beta^T \ln x_i + \varepsilon_i, \qquad \varepsilon_i = v_i - u_i \qquad (1\text{-}17)$$

In this case let $u \sim N(0,\sigma_u^2)$ truncated below at 0[10]. (Note that σ_u^2 is the variance of the underlying normal distribution. The variance of u will be denoted by Var(u).) The distribution function of ε is the distribution of the sum of a symmetric normal random variable and a truncated normal random variable. (Weinstein (1964) see Appendix 1.)

$$h(\varepsilon) = \frac{2}{\sigma} f^*\left(\frac{\varepsilon}{\sigma}\right) F^*\left(\frac{-\varepsilon\lambda}{\sigma}\right) \qquad (1\text{-}18)$$

[10] Other distributions for the one sided error term, include exponential, half-normal with non-zero mode, log-normal, gamma - see Stevenson (1980), Greene (1980), Greene (1990) (Appendix 1).

where $\sigma^2 = \sigma_u^2 + \sigma_v^2$ and $\lambda = \frac{\sigma_u}{\sigma_v}$.

The mean and variance of this distribution are given by

$$E(\varepsilon) = E(u) = \frac{\sqrt{2}}{\sqrt{\pi}}\sigma_u \quad \text{for proof see Appendix 1.2.1} \qquad (1\text{-}19)$$

$$\begin{aligned} Var(\varepsilon) &= Var(u) + Var(v) \\ &= \left(\frac{\pi - 2}{\pi}\right)\sigma_u^2 + \sigma_v^2 \quad \text{for proof see Appendix 1.2.1} \end{aligned} \qquad (1\text{-}20)$$

Once the stochastic frontier has been estimated using maximum likelihood estimation or COLS (see Appendix 1), the average efficiency for the industry can be determined. The technical inefficiency, u_i, for each observation is required.

1.3.3 Separating the error term into two components

To estimate the firm specific efficiencies the conditional distribution of u_i given ε_i is used. The mean or the mode of this distribution gives a point estimate of u_i.

The conditional distribution of u given ε, where u is half-normal, is that of a $N(\mu^*,\sigma^{*2})$ variable truncated at zero (Jondrow et al. (1982)) (where $\mu^* = \frac{-\sigma_u^2 \varepsilon}{\sigma^2}$, $\sigma^* = \frac{\sigma_u^2 \sigma_v^2}{\sigma^2}$ and $\sigma^2 = \sigma_u^2 + \sigma_v^2$).

The mean of this distribution is given by

$$E(u|\varepsilon) = \mu^* + \sigma^* \frac{f^*\left(\frac{-\mu^*}{\sigma^*}\right)}{1 - F^*\left(\frac{-\mu^*}{\sigma^*}\right)} \qquad (1\text{-}21)$$

where f* and F* are the standard normal density and distribution functions respectively.

Equivalently

$$E(u|\varepsilon) = \sigma^* \left[\frac{f^*\left(\frac{\varepsilon\lambda}{\sigma}\right)}{1 - F^*\left(\frac{\varepsilon\lambda}{\sigma}\right)} - \left(\frac{\varepsilon\lambda}{\sigma}\right) \right] \qquad (1\text{-}22)$$

where $\lambda=\frac{\sigma_u}{\sigma_v}$.[11]

The mode M(u|ε) can also be found - it is the minimum of μ* and zero. Either of these can be used to give a measure of the efficiency for each unit. Throughout the thesis E(u|ε) will be used as the SF efficiency estimate of an individual DMU.

The error term has now been separated into two components for each observation, so a measure of the efficiency (dependent on the total error) for each firm can be found and confidence intervals for u can be computed.

Throughout the thesis the stochastic frontier will be estimated by maximum likelihood using the LIMDEP software (Greene[12]). This method generally gives better estimates than COLS methods (it starts from the COLS estimates) and is less likely to involve variance

[11] λ is a measure of the relative variability of the two sources of random error.
$\lambda^2 \rightarrow 0 \Rightarrow$ The density function of ε becomes the density of a $N(0,\sigma^2)$ random variable.
$\lambda^2 \rightarrow \infty \Rightarrow$ The density function of ε becomes the density of a negative half-normal random variable.

[12] Note that there are some mistakes in the description of the SF method in the LIMDEP v6 and v7 reference manuals.

decomposition errors[13] as encountered by almost all of the data sets in Banker et al. (1993). These errors occur when the ratio of the variance of the inefficiency term to the variance of the random error term is high or low: Olson, Schmidt and Waldman (1980);

> "As $\lambda \rightarrow 0$ ($\sigma_u^2 \rightarrow 0$) the probability of a Type I failure approaches (approximately) 1/2 ... Type II failures occur with non-negligible probability when λ is large ... There appeared to be no comparable problem with MLE."

Several different functional forms have been proposed to represent the production technology. The simplest, and easiest to use is the Cobb-Douglas function (Cobb and Douglas (1928)). Other more flexible forms are also used e.g., the Constant Elasticity of Substitution (CES) (Arrow et al. (1961), Leontief, Generalised Leontief (Diewert (1971)), translog (Christiansen, Jorgensen and Lau (1971)), etc. Gong and Sickles (1992) used a Monte-Carlo analysis to investigate the benefits and drawbacks of some of these forms for efficiency measurement (this paper will be discussed in more detail in the next chapter). See also Guilkey, Lovell and Sickles (1983).

[13] Type I failures occur when the variance of the inefficiency is computed to be negative. Type II failures occur when the variance of the measurement error is computed to be negative.

A Fourier flexible form[14] has been proposed as a more flexible form than any of the functional forms mentioned above (Gallant (1981) and (1982).) The Fourier form is said to be more flexible than the translog as a translog function is a Taylor expansion about a point. This means that it is only accurate locally. Mitchell and Onvural (1996);

> "The Translog represents a second-order Taylor series approximation of an arbitrary function at a point. ...least squares estimates of a second-order polynomial such as the Translog do not generally correspond to the Taylor series expansion of the underlying function at an expansion point and, hence, are biased estimates of the series expansion."

The problem with applying Fourier series in practice is that in order to specify the function, a large number of terms are needed, even in the case with only 2 independent variables. In the examples shown later in the thesis, the translog form has been used when a flexible functional form is required, as in each case it is shown to be flexible enough to capture the underlying nature of the technology. However, with more

[14] Mitchell and Onvural (1996): "It is well known from advanced calculus that a linear combination of sine and cosine functions called a Fourier series can represent exactly any well-behaved multivariate function f(x). This is possible because sine and cosine functions are mutually orthogonal and function space-spanning; hence, representing an arbitrary function by a Fourier series is analogous to representing an n-vector as a linear combination of n mutually orthogonal, function space-spanning basis vectors. Thus a researcher lacking knowledge of the true form of a cost function may avoid gross functional misspecification by positing a Fourier series."

complex technologies the Fourier form may indeed give better estimates and could be used if functional misspecification is suspected in the translog specification.

1.4 Advantages and disadvantages of the methods

1.4.1 Advantages of DEA

DEA is non-parametric, which means that the danger of imposing the wrong functional form is avoided. DEA makes few assumptions about the form of the technology. (See Chapter 2 for further discussion of the assumptions.) SF on the other hand, requires assumptions about the form of the relationship between the inputs and outputs and the distribution of the random error and inefficiency terms. A parametric function has the disadvantage that it is restricting but it is useful to be able to characterise the production technology in a simple mathematical form.

DEA can easily handle multiple inputs and outputs as opposed to the usual stochastic frontier formulation, which is restricted to the single output case when estimating a production technology. However, it is now possible to include multiple outputs in a parametric analysis - see Chapter 6.

DEA also readily gives extra information in the form of peers and targets. The peer units of a particular DMU_0 are those efficient units that DMU_0 is being compared to. These are generally units operating at a similar scale size and mix to DMU_0, enabling the decision-makers at DMU_0 to compare itself to similar units that are performing better. The targets are the values of the inputs and outputs that the DMU should be able to achieve once it becomes efficient.

1.4.2 Advantages of SF

SF allows random noise to be incorporated into the model. DEA is deterministic which means that it assumes that there is no random noise in the data - this is a very big assumption: Any statistical noise, measurement errors, luck, omitted variables and other misspecifications are counted as inefficiency. Deterministic models are very sensitive to outliers. We will see just how much this can affect the relative performance of the methods in Chapter 3.

Because SF is a parametric method based on regression, it is possible to create confidence intervals for the parameters in a SF model. DEA is non-parametric but the statistical properties of the efficiency

estimates are now being developed (see Banker (1993) and Simar (1997)[15]).

When dealing with panel data, i.e. data over several time periods, the SF method has the added advantage that it is no longer necessary to specify the distribution of the inefficiency term. However, panel data sets will not be considered here.

1.5 Summary of Chapter 1

In this chapter, an overview has been given of the theory of production as it relates to the measurement of technical efficiency. The different measures of technical efficiency (pure technical efficiency and scale efficiency) have been described along with the concept of returns to scale. The stochastic frontier and DEA methods have been outlined along with their comparative advantages and disadvantages.

In the next chapter the possible differences between the estimates of the methods are highlighted and hypotheses are put forward as to the

[15] Banker (1993) showed for the single input, multiple output case that the DEA estimates are consistent. Park, Kneip and Simar (1996) showed that the DEA estimates are consistent for multiple inputs and outputs.

effects on the methods of assumptions of either method not being met by the underlying data.

Chapter 2

Comparing the estimates of DEA and SF

2.1 Introduction

Consider a set of DMUs with certain inputs and outputs. For given inputs we say that there is some set of maximum output values which the DMU could achieve, assuming that the output mix remains the same. These output values are given by the true production frontier. Any efficiency estimation method will estimate this frontier by some process and use the distance of the unit from the estimated frontier as a measure of the inefficiency of the unit. Because the underlying natures of the DEA and SF methods are very different it is likely that there will be some differences between the estimates obtained from these methods. Where the methods give different estimates it is necessary to be able to identify why there is a difference and which method is giving the better estimates.

There are three possible observations when two estimating methods are employed: they give the same estimates; the first method gives larger estimates than the second; or the first gives smaller estimates than the second. For the latter two cases we need to be able to say which of the two methods is giving better estimates. To do this, the underlying nature of the methods will be drawn upon.

> "Consider ... a comparison of an econometric flexible form estimation and a non-parametric estimation of a frontier; such a comparison will highlight

differences according to the premises behind the choice of models within this comparison." Olesen (1995).

This chapter will identify the premises of each method and give hypotheses as to how the violation of these will affect the results of the methods.

The next section gives a brief discussion of three previous comparisons of the methods. Section 2.3 discusses the implicit assumptions that each of the methods makes. Section 2.4 discusses a measure which is used throughout the thesis to examine the 'goodness of fit' of the estimated frontier to the true frontier and Section 2.5 examines the possible differences that can be observed between the efficiency estimates of the two methods and how the estimates may relate to the true values. Finally, Sections 2.6 and 2.7 put forward the hypotheses that will be investigated and discussed in the following chapters of the thesis, and Section 2.8 concludes.

2.2 Previous comparisons of the methods

Although both DEA and SF methods have been used for several years now, there has been no systematic comparison of the two methods and there is still relatively little literature on their comparative performance. Banker and Cooper (1994) summarise three comparisons that have used simulated data.

The first, Banker et al. (1988) compared DEA with a deterministic frontier. The main conclusion of this paper was that the observations that are likely to be misclassified by DEA are the 'corner' points - i.e. units with a very small or very large value for at least one of the inputs or outputs. The DEA results were found to be much better than the translog deterministic frontier estimates. However, the true frontier was piecewise log-linear (i.e. not continuously differentiable) which is an advantage to DEA, and the underlying inefficiency distribution was uniform. Chapter 3 will investigate the effect of the underlying inefficiency distribution on the frontier methods.

The deterministic frontier in Banker et al. (1988) was expanded to a stochastic frontier in Banker, Gadh and Gorr (1993). However, in this case the frontier was estimated using COLS which gave a very large number of type I and type II failures (see footnote 13 in Chapter 1 for definitions). The authors found that DEA gave closer estimates[1] to the true values than COLS for small samples and low random noise. For large samples (N > 100) and high random noise, COLS gave results that were "about 25% more accurate than DEA". For low levels of random noise, the further the inefficiency distribution was from the assumed half-normal distribution, the more accurate were the DEA results in comparison with those of COLS. The authors concluded that

[1] measured with Mean Absolute Deviations (MADs) and mean deviations.

both DEA and COLS performed unacceptably in the presence of high random noise. They found that if the better method were chosen in each case, the MADs for low random noise were about 0.05 to 0.06. If the worse method were chosen, the MADs were in the region of 0.09 to 0.10. Therefore, a significant improvement could be made if the better method could be identified. It is this identification, using a comparison between the results of the methods, which is the aim of this thesis.

The other simulation comparison discussed by Banker and Cooper was Gong and Sickles (1992). This paper compared DEA and SF using simulated panel data and cost frontiers rather than production frontiers. In this paper the authors found that the choice of functional form in the SF method is important and that when there are correlations between the inefficiency term and the inputs, the SF method is adversely affected. Other conclusions in this paper related to the panel nature of the data, which will not be considered here.

Since the Banker and Cooper (1994) paper, Banker, Chang and Cooper (1996) published a further simulation study. Once again, this paper compared DEA with a deterministic frontier using COLS as in Banker et al. (1988). This paper compared the relative performances of the methods for different sample sizes and also investigated the effects of omitted/irrelevant variables and collinearity.

All of these papers involved Monte-Carlo comparisons between the two methods and the performance of each method was measured using the mean absolute deviation (or mean deviation) of the estimates from the true efficiency value or the correlation between the estimates and the true values. All these measures are average measures across the whole technology. In this thesis it is proposed that, by looking at how the estimates vary across specific dimensions of the technology, it may be possible to identify which of the methods is giving the better estimates in specific regions of the technology by identifying which of the underlying assumptions of the methods have been violated.

In Chapter 8, an algorithm will be developed for comparing the results from the two methods. The combination of applying both methods and comparing the results leads to a way of validating the results from either method and enables the analyst to choose the best model specification.

Note that Arnold et al. (1996) also proposed a combination of DEA and SF approaches. This method first uses the DEA results to split the DMUs into those that are efficient and those that are inefficient and then applies a regression method to the two sets. However, this method still assumes that the correct DEA model has been used initially. The algorithm developed in this thesis should be able to detect which of the DEA models should be used.

2.3 Assumptions

An efficiency estimation method makes certain assumptions about the true technology. If these assumptions are valid then the method is likely to give efficiency estimates which are very close to the true values. However, if some of the assumptions do not hold true, then there will be regions of the technology where the estimates will be very far from the true values. A region may consist of the whole technology or it may be a very small subset of the technology.

Definition: A **region** is a set of feasible input-output combinations which fall within a certain range of scale sizes or input mixes.

When there are regions of good and regions of poor estimation, we say that there is '**variation of fit**' of the estimating method to the true technology. This concept will be described fully in Section 2.4 of this chapter.

2.3.1 Assumptions of Data Envelopment Analysis

The PPS formed by the DEA postulates (Olesen (1995)) relies on certain more general assumptions about the nature of the underlying data:

DEA A1. No random noise.

DEA A2. CRS (in some models this assumption is relaxed).

DEA A3. There is a good spread of efficient units across the whole technology.

DEA A4. Convexity of the production possibility space.

The effects of the first three of these assumptions will be examined in later chapters of the thesis and the fourth assumption will be discussed in Section 2.7 of this chapter.

2.3.2 Assumptions of Stochastic Frontiers

Stochastic frontiers impose behavioural assumptions on the production function and distributional assumptions on the error term:

SF A1. Distribution of the random noise term.

SF A2. Distribution of the inefficiency term.

SF A3. Form of the production function (CRS, CES, etc.).

SF A4. No correlation between the inefficiency and the exogenous variables.

SF A5. Inefficiency only in endogenous variable.

The first three of these assumptions are examined in the thesis. The last two assumptions are discussed in Section 2.7 of this chapter.

Note that if any other assumptions are made which have not been detailed above, the conclusions made in Chapter 8 may need

modifying. For example, it could be proposed that another assumption made by both methods is that all necessary variables have been included in the analysis, which will affect the results in both approaches. However, rather than considering this as a separate assumption, here it is assumed that omitting variables will have a similar effect to increasing the level of random noise. (See Banker et al. (1996).)

The performance of each method relies upon the validity of the assumptions it makes.

It may happen that one or more of these assumptions does not hold. This could lead to the method performing poorly across the whole technology - as in the case of the assumption of no random noise not holding for DEA - or the assumption may be valid in certain regions of the technology and not others. In this case the method will give good estimates in certain regions of the technology and not in others. This will lead to a variation in the fit of the estimating frontier to the true frontier.

2.4 Variation of Fit

When either DEA or SF is applied to a data set, the estimated efficiencies are likely to be close to the true efficiencies for some DMUs and further away from the true values for other DMUs. This variation in

the proximity of the estimating and true functions will be termed **variation of fit**.

2.4.1 Measuring Variation of Fit

In order to measure the variation of fit, the amount that the estimated function under SF deviates from the true function across the technology is examined. This measure is termed the **Functional Deviation** (FD) and is defined, for the output orientation, for DMU j as:

$$FD_j = \frac{\text{estimated efficient output of DMU j}}{\text{true efficient output of DMU j}}. \qquad (2\text{-}1)$$

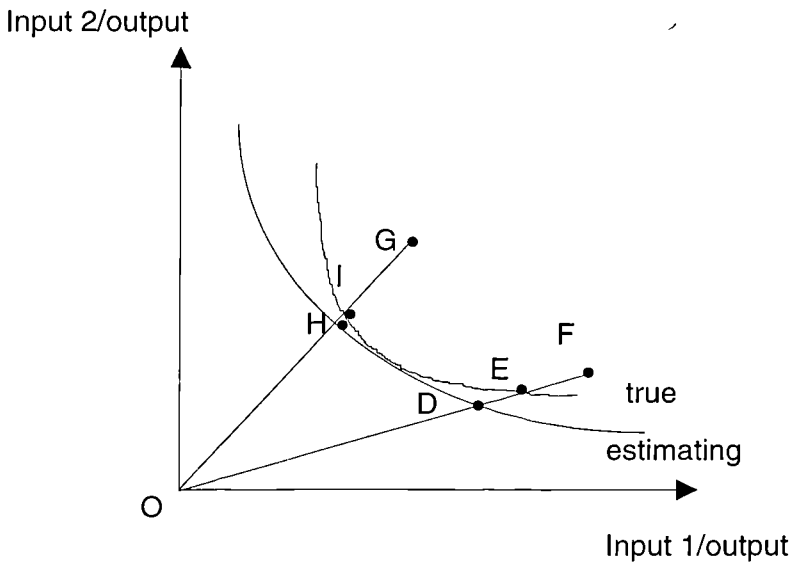
For the input orientation, the FD is defined as:

$$FD_j = \frac{\text{true efficient input of DMU j}}{\text{estimated efficient input of DMU j}}. \qquad (2\text{-}2)$$

Note that these are equivalent for the case of a CRS frontier.

The estimated efficient output (choose a particular output in the case of multiple outputs) or input is taken from the core function - this is the level of output which the firm could achieve at its current input mix, if there was no inefficiency and no random noise. Thus the random noise is ignored and only the differences between the true and estimating functions are considered. The FD will be close to 1 when the

Figure 2-1. Functional deviation on input mix



estimating function is very close to the true function. As the estimated function diverges from the true function, the value of the functional deviation moves away from 1.

When there is no random noise in the data, the functional deviation can also be written as

$$FD_j = \frac{\text{true output efficiency of DMU j}}{\text{estimated output efficiency of DMU j}}. \qquad (2\text{-}3)$$

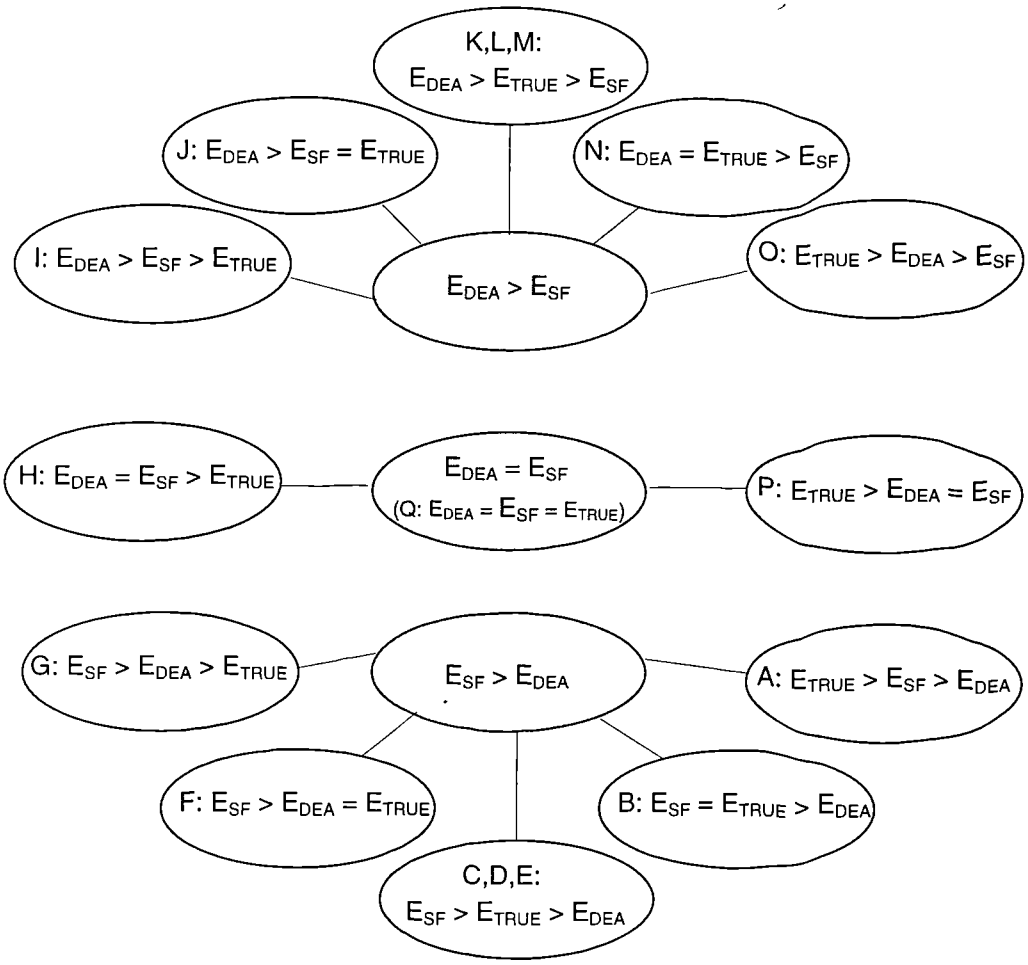
The definition of functional deviation is illustrated using Figure 2-1 (note that this technology has CRS).

For DMU F, $FD_F = \frac{OE}{OD}$ while for DMU G, $FD_G = \frac{OI}{OH}$. Clearly $|1-FD_G| < |1-FD_F|$ and there is a difference of fit between the estimated efficient projections of F and G to the true projections. In this case of constant returns to scale, obviously any unit with the same input mix will have the same functional deviation as the FD measures the distance between the true and the estimated frontier.

2.5 Possible differences between DEA and SF efficiency estimates

In Figure 2-2 all the possibilities are given for the efficiency estimates for an individual DMU under SF and DEA. (The letters in Figure 2-2 relate to the regions shown in Figure 2-3 when Method 1 refers to DEA and Method 2 to SF.) There are three possible outcomes for each DMU; the efficiency estimates under both methods are equal; the estimate under DEA is greater than the estimate under SF; the estimate under SF is greater than the estimate under DEA. These are the only possible outcomes. However, the true efficiency value may lie between the estimates or beyond either of them.

Figure 2-2. Possible differences between DEA and SF estimates



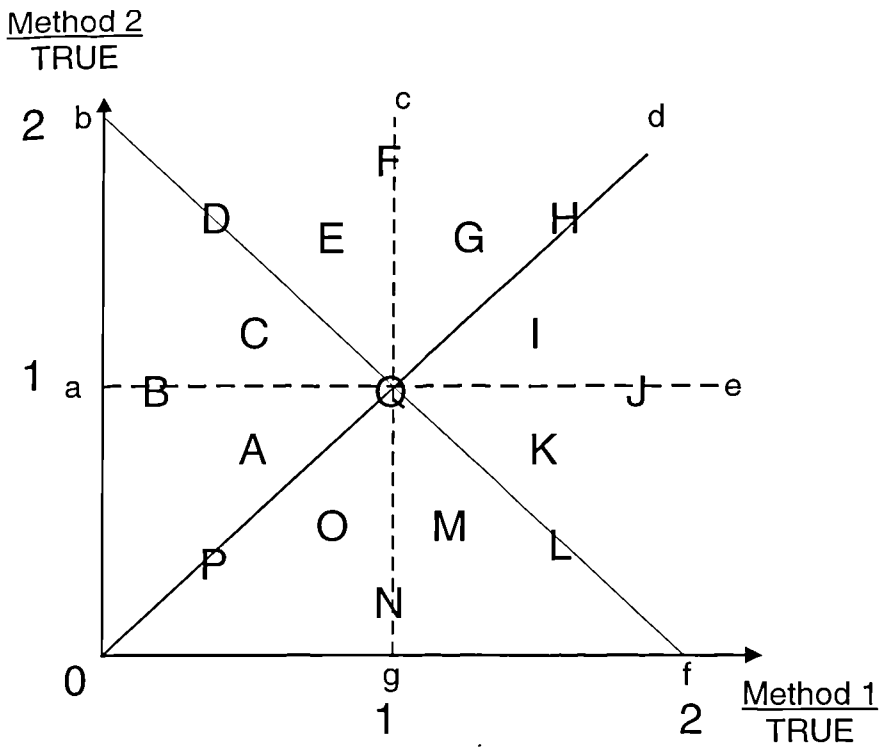
Whenever the methods do not give similar estimates, the decision-maker needs to be able to tell which of the methods is giving closer estimates to the true values.

In Figure 2-3 the ratios of the estimated efficiency to the true efficiency of two methods are plotted against each other. These methods could be, for example, DEA and SF, or two SF methods utilising different

inefficiency distribution assumptions, or two DEA methods having different assumptions about the returns to scale.

When the efficiency estimated by Method 1 is equal to the true efficiency, the DMU will lie on the line **cg** and when the efficiency estimated by Method 2 is equal to the true efficiency, the DMU will lie on the line **ae**. Therefore, the only point on the graph where both methods give good estimates of the true efficiency is the intersection of **cg** and **ae**, point Q.

Figure 2-3. Comparing the estimates from two methods



Let us take the case where Method 1 is DEA with certain assumptions and Method 2 is SF with certain other assumptions.

In a region of the technology the SF estimated efficiencies are found to be greater than the DEA estimates. The DMUs in this region may lie in sections A, B, C, D, E, F or G in Figure 2-3.

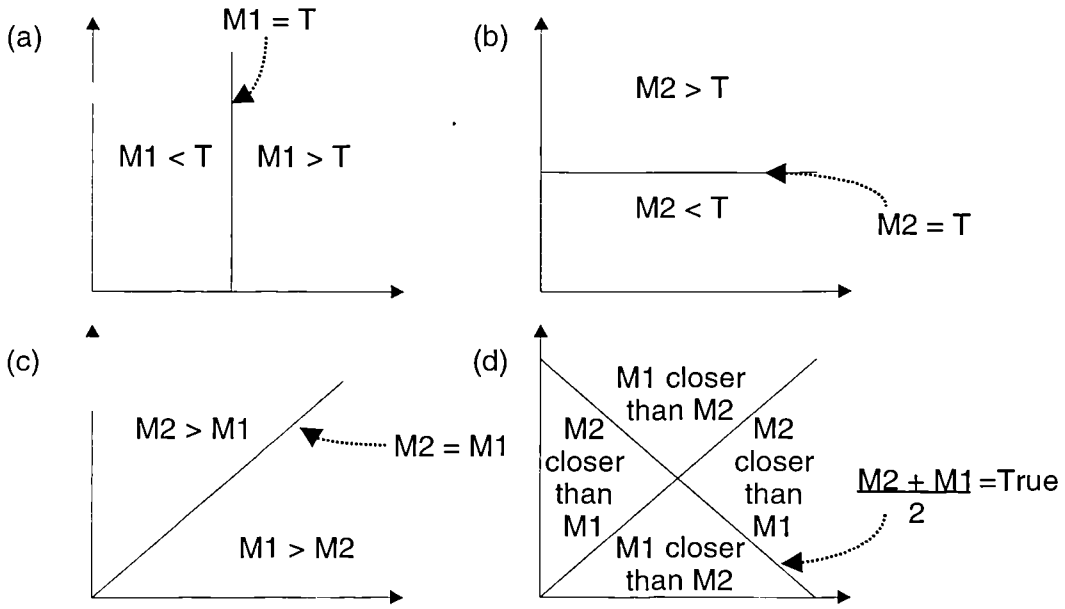
- In section A, $E_{TRUE} > E_{SF} > E_{DEA}$
- In section B, $E_{TRUE} = E_{SF} > E_{DEA}$
- In sections C, D and E, $E_{SF} > E_{TRUE} > E_{DEA}$
- In section F, $E_{SF} > E_{TRUE} = E_{DEA}$
- In section G, $E_{SF} > E_{DEA} > E_{TRUE}$

To be able to identify which of the methods is giving estimates that are closer to the true values, we need to decide whether the DMU lies in A, B or C, in which case SF is giving closer estimates than DEA; or in D where the true value lies exactly between the two estimates; or in E, F or G, in which case the DEA estimates are closer than SF.

This is a very useful diagram for comparing how the methods are performing on different data sets when we know the true technology as we can clearly see across the whole technology whether there are any units in each of the regions. It will give a clear picture of how much an assumption affects the performance of the methods as it can easily be compared for different data sets.

In Figure 2-4 the outline of the graph in Figure 2-3 is redrawn highlighting different aspects of the graph. In Figure 2-4(a) we can see that the line **cg** in Figure 2-3 divides the graph between the areas where Method 1 is giving lower estimates than the true values and areas where Method 1 is giving higher estimates than the true values. For points on the line **cg**, the estimates of Method 1 are equal to the true values.

Figure 2-4. Aspects of Figure 2-3



Similarly, diagram (b) shows that the line **ae** cuts the graph into two regions, one where Method 2 gives smaller estimates than the true values and one where it gives larger estimates than the true values.

Diagram (c) shows that the 45° line through the origin, **0d**, separates the units into those for which Method 2 is giving greater estimates than Method 1 and vice versa. The line **bf** gives all the points where the estimates from both methods are equally far from the true values. Diagram (d) shows how this line and the line 0d separate the units into four regions: two regions where Method 1 is closer than Method 2 and two regions where Method 2 is closer than Method 1.

The point Q, in the centre of the graph is obviously the only point where both methods give equal estimates and where the estimates are equal to the true values. Therefore, we would like the points to be as close as possible to the centre point on the graph - i.e. the results from both methods being very accurate - or very close to either the horizontal or vertical lines **ae** or **cg** - i.e. one of the methods giving very close estimates to the true values.

In a real application, all that one can say is whether the estimates given by Method 1 are larger than or smaller than those given by Method 2. We need to be able to say, by reasoning from the underlying nature of the methods, which of the areas the DMU is most likely to be lying in so that we can deduce which of the Methods is giving us the best estimate of efficiency.

When one assumption at a time is violated and all the rest hold true, it may be possible to say which regions the DMUs are most likely to be in. A number of alternative scenarios will be presented of differences between the estimates and the true efficiency values for different underlying potential causes. These causes will be violations of various assumptions underlying DEA and SF. The assumptions were listed earlier.

2.6 The Hypotheses which will be investigated in the thesis

2.6.1 Differences across the whole technology

Some of the assumptions listed in Section 2.3 will affect units lying in any region of the technology rather than affecting units because of their size of operation, or mix of inputs or outputs. These are the assumptions which will be investigated in Chapter 3.

2.6.1.1 DEA A1 does not hold: The data contains random noise

DEA assumes that there is no measurement error in the data and attributes all deviations from the estimated frontier to inefficiency.

If the data does contain random noise, then for an output orientation, the observed values of the outputs will be greater than or less than the true output values. The units that are given lower values than the true

values will not affect the placement of the production frontier unless they happen to lie on the true frontier before the noise is added. However, the units which are given higher values than the true values will move up, towards the frontier, and any units which were on the true frontier or just below it, will push the observed frontier upwards. If the assumption that the random noise is symmetrically distributed holds, there are equally many units that move up as move down, so it is likely with a large enough data set that the observed frontier lies above the true frontier whenever there is any random noise. Of course, there is a small possibility that the frontier could move down if all the units on, or just below, the frontier have negative random noise, but as the sample size increases this becomes increasingly improbable.

Therefore, DEA is expected to give lower efficiency estimates on average than the true values whenever there is random noise, because the noise will push some of the units above the true frontier effectively shifting the whole frontier up.

The SF method allows for random noise so the estimates given by the SF method should be close to the true efficiencies. (This assumption will be discussed in Chapter 3.)

Let $\overline{E}_{TRUE}$ denote the average efficiency of all the units in the data set and $\overline{E}_{SF}$ and $\overline{E}_{DEA}$, the average estimates under SF and DEA respectively. Then:

Hypothesis 1 (illustrated in Chapter 3)

If the data contains random noise then the whole technology will be given estimates such that $\overline{E}_{TRUE} \cong \overline{E}_{SF} > \overline{E}_{DEA}$.

2.6.1.2 SF A1 does not hold: The random noise term is not normally distributed

The random noise term is always assumed to be normally distributed as it is attributed to measurement error, luck, etc., which should have symmetric effects.

However, in the SF method the random noise term is always assumed to be additive, i.e. the specification of the SF model is

$$y = f(\beta;x) + v - u. \qquad (2\text{-}4)$$

In the case of a Cobb-Douglas function this model becomes

$$\ln(y) = \ln(A) + \sum_i \beta_i \ln(x_i) + v - u. \qquad (2\text{-}5)$$

The random noise term in this case is v. However, the frontier could be specified equally well as

$$y = A \prod_i x_i^{\beta_i} e^{v-u}. \qquad (2\text{-}6)$$

In this case, the random noise is multiplicative and equal to e^v. If $e^v \sim N(0,\sigma_v^2)$ then $v \sim \ln N(0,\sigma_v^2)$.

If the random noise is assumed to be additive in all cases, and normally distributed at the raw data level then the frontier will be given by

$$y = A \prod_i x_i^{\beta_i} e^{-u} + w \qquad (2\text{-}7)$$

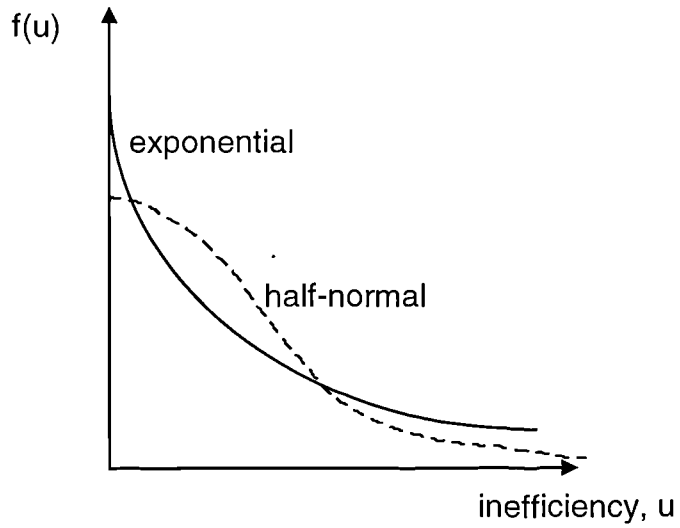
where w is now the random noise term, $w \sim N(0,\sigma_w^2)$.

In each of the cases given by equations (2-5), (2-6) and (2-7), the random noise in the log-linear specification has a different distribution than the raw data level. In each case, at one of the two levels, the random noise will depend on the level of output (i.e. is heteroskedastic). This problem will be discussed further in Chapter 3.

2.6.1.3 SF A2 does not hold: How dependent is the SF method on the assumption made about the inefficiency distribution?

Suppose that the half-normal is the true inefficiency distribution and an exponential is used as the estimating distribution for the inefficiency term (see Figure 2-5).

Figure 2-5. Possible inefficiency distributions



There will clearly be differences between the estimated and true values across the true inefficiency values.

If the inefficiencies were estimated independently of the total error then a comparison between the SF and DEA estimates should be able to identify possible misspecification of the inefficiency distribution. The DEA estimates do not depend on a distributional assumption so any deviations between the DEA estimates and the true values should not

depend on the level of the efficiency. Therefore a comparison between the SF estimates and the DEA estimates across the DEA estimates should be able to identify whether the SF estimates depend on the level of true efficiency.

However, the inefficiency cannot be separated completely from the total error term; the values obtained from the SF method are still dependent upon the total error.

The variance of the total error term can be split into the variance of the inefficiency term plus the variance of the random noise. If an incorrect assumption is made about the inefficiency distribution in the SF method then the difference between the estimated distribution and the true distribution will be attributed to random noise. So the estimated variance of the random noise term will be greater than the true value and the variance of the inefficiency term will be less than its true value.

If an incorrect inefficiency distribution is imposed, the SF method compensates by adjusting the level of random noise assumed in the data.

This leads to:

Hypothesis 2 (illustrated in Chapter 3)

If the assumed inefficiency distribution is not the same as the true inefficiency distribution, the SF method will compensate by allocating more of the total error to random noise. The differences between the true and estimated efficiencies will vary as the true inefficiency changes.

2.6.2 Differences across scale size

Some of the assumptions of the methods will affect units differently according to their scale of operation. These assumptions will be illustrated in Chapter 5. (Chapters 4 and 6 will define the measure of scale that will be used to separate the technology into different regions of scale size.)

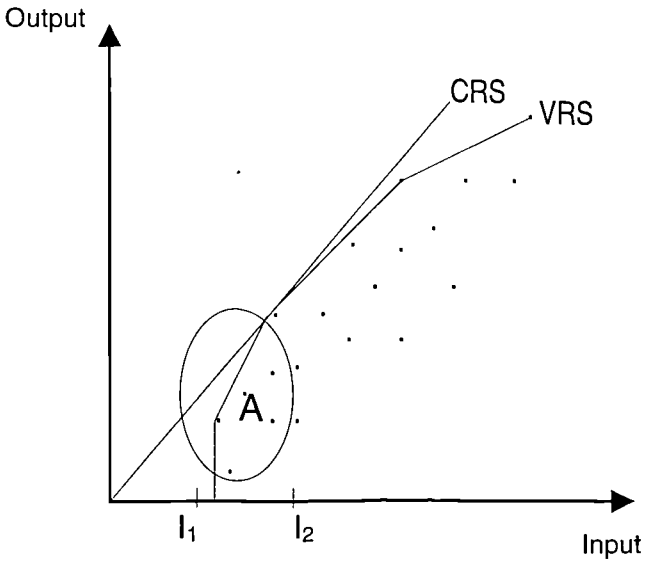
2.6.2.1 DEA A2 is not valid: A too restrictive assumption about returns to scale is imposed

Consider the graph shown in Figure 2-6 for a single-input, single-output technology. The true technology is given by the line labelled VRS. If the CRS DEA method is imposed on this technology the frontier given by the line labelled CRS will be used as the benchmark against which the DMUs are measured. Obviously, the efficiencies in areas where constant, or nearly constant, returns to scale hold are likely to be well estimated by both the SF and DEA methods. However, in other areas,

the efficiencies will be under-estimated, as the estimated function is too restricting.

For example, in region A in Figure 2-6, the units will all be given efficiency estimates which are too low due to scale inefficiency in this region.

Figure 2-6. Differences in specification under constant and variable returns to scale



Note that there can be some confusion when discussing assumptions about the nature of returns to scale on the frontier. The CRS assumption is the most restrictive assumption regarding returns to scale but is given by the DEA model with the least constraints. NIRS and

NDRS are both less restrictive than the CRS assumption and VRS is the least restrictive assumption.[2]

Let E_{DEA} denote the efficiency estimated under DEA of a particular DMU, E_{SF} the efficiency estimated under SF and E_{TRUE}, the true efficiency of the same DMU.

Hypothesis 3 (illustrated in Chapter 5)

If a restrictive assumption of returns to scale is imposed on the methods unnecessarily, then the estimates will be such that $E_{TRUE} > E_{DEA}$ in the regions where the assumption does not hold. For an homothetic[3] frontier, these regions will vary *only* across scale size.

Similarly, by assuming a functional form which has CRS in the SF method, the CRS assumption can be incorrectly imposed. This will lead to functional misspecification in the SF method - see Section 2.6.3.2.

---

[2] See Seiford and Thrall (1990) for a discussion of these models and their assumptions.

[3] See Chapter 6 for a discussion of homotheticity.

2.6.2.2 DEA A2 does hold but this assumption is relaxed by the method

If the majority of the DMUs operate at a certain scale size (this is very likely in many industries) and the few DMUs at other scale sizes are inefficient then not imposing a CRS frontier on a CRS data set will cause problems.

In the case of a flexible stochastic frontier function, the areas where there are very few DMUs which are all inefficient, are likely to 'pull' the function towards the DMUs and make these DMUs appear more efficient than they actually are.

In the case of DEA, the frontier will be heavily influenced by exactly where the efficient units are. Once again, consider the graph in Figure 2-6. If the CRS frontier is now the true frontier and the VRS frontier is used in a DEA method, the DMUs in region A will once again be given poor efficiency estimates, but in this case they will be given overestimates of the true efficiencies.

The inefficient cluster of DMUs at A leads to bad specification of the true frontier for inputs in the range I_1 - I_2. This misspecification could occur locally in any region where there is a cluster of inefficient DMUs.

Hypothesis 4 (illustrated in Chapter 5)

If the true technology has CRS, NIRS or NDRS and a less restrictive assumption is imposed on the estimating methods then the estimates will be such that $E_{DEA} > E_{TRUE}$ in the regions where the assumption does not hold. These regions will vary across scale size.

2.6.3 Differences which may occur across scale size or input mix

The next two hypotheses relate to assumptions which may affect units according to their scale size or their input mix.

2.6.3.1 DEA A3 does not hold: There is not a good spread of efficient units across the whole technology

In many real data sets it is likely that most of the DMUs have similar operating mixes and similar scale sizes. Away from this main scale size and mix, there will be fewer DMUs. As the DEA method is very dependent upon where the efficient DMUs are, in order to form the frontier, the method not only needs a large number of units in order to give a good estimate of the frontier, but also needs the efficient units well spread across the technology.

Once the data set has regions where there are few units, it becomes more likely that these regions will not include an efficient unit - leading to the technology in that region not being well specified by DEA. If this region is in the centre of the technology and the convexity assumption

holds, this is not a problem as the surrounding regions should still be able to define the frontier in this region. However, if the region of scarcity is at the 'edge'[4] of the technology there are more problems.

These problems have been addressed above in section 2.4.3 for a scarcity of units at extreme scale sizes. When the region of scarcity is at an extreme input mix there will be similar problems. The estimated technology will be closer to the units than the true technology because of the lack of efficient units to form the frontier. This will lead to the DEA estimates being greater than the true efficiency values.

Hypothesis 5 (illustrated in Chapter 8)

If the technology has few units in certain regions of input or output mix, and these regions are at the edge of the technology, then the DMUs in these regions may be given estimates such that $E_{DEA} > E_{TRUE}$.

2.6.3.2 SF A3 does not hold: The true technology is not well specified by the estimating SF function

There are several possibilities here. The cases which will be investigated are when

[4] The term 'edge' applies to units which have a large or small value for one of the variables (compared to other units in the data set). If one of the input values is large or small, the unit will be at an extreme input mix. On the other hand if the output value is large or small (in the single output case) the unit will be at an extreme scale size.

- the true function is piecewise log-linear (i.e. not a continuously differentiable function) - See Chapter 4;
- a too restrictive assumption of the returns to scale is imposed - See Chapter 5;
- the true function has a low or high elasticity of substitution and the estimating function does not pick this up - See Chapter 7.

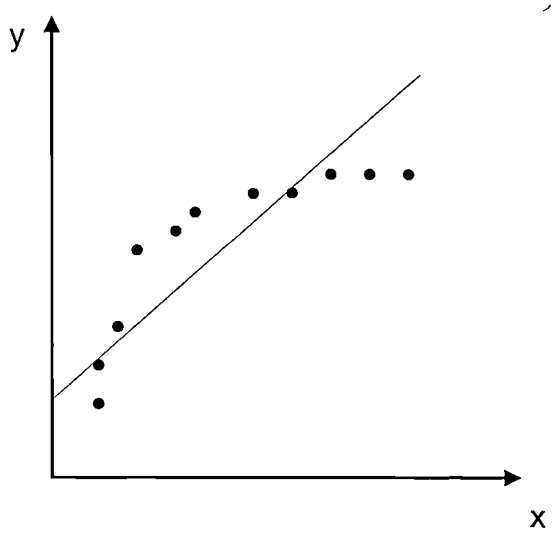
Any other case where the estimating function imposes restrictions on the form of the technology could lead to functional misspecification, e.g. imposing homotheticity[5] on a non-homothetic technology.

In any of these cases, if there is variation in the fit of the estimated function to the true technology due to the restrictions imposed by the estimating function, there will be definite regions where the SF method gives estimates which are greater than the true efficiency values and other regions where the SF estimates are less than the true efficiency values.

For example (see Figure 2-7), in the case of OLS regression, if a linear function is imposed on a quadratic relationship, there will be definite regions where the estimated y value is less the true value and other regions where it is greater.

[5] See Chapter 6.

Figure 2-7. Ordinary least squares regression of a linear function



How these regions vary across the technology will depend on the restriction that is imposed. If the restriction is on the returns to scale of the frontier then the differences between the estimated efficiencies and the true efficiencies will vary across scale size. However, if the restriction is on the elasticity of substitution, the regions will vary across the input mix.

Hypothesis 6 (illustrated in Chapters 3, 5 and 7)

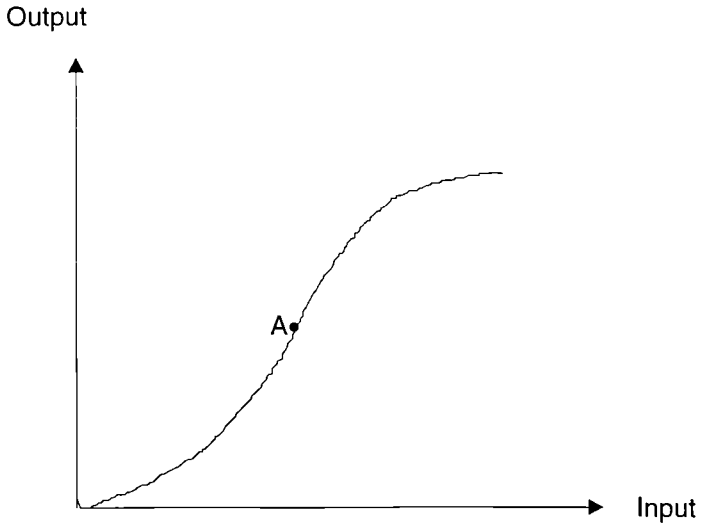
If the true frontier is not well estimated by the SF function, then the estimated efficiencies will have regions where they are greater than and less than the true efficiencies across scale size or input mix, depending on whether the misspecification varies across scale size or input mix.

2.7 The hypotheses which will not be investigated in this thesis

2.7.1 DEA A4 does not hold: The true technology is non-convex[6]

In the economic production theory literature there is no assumption of convexity. The basic assumption is that no output can be produced from no input (Shephard (1970)), implying that the production function must always go through the point ($\mathbf{x} = \mathbf{0}$, $\mathbf{y} = \mathbf{0}$). This leads to production being given by an S-shaped curve (Figure 2-8) - i.e. at input levels of zero, no output can be achieved; as the input increases, the technology exhibits increasing, constant and finally decreasing returns to scale.

Figure 2-8. An S-shaped curve



[6] Note that it is assumed throughout the thesis that the input and output sets are convex.

If convexity of the whole production possibility space is imposed as in DEA, only a concave production function is allowed, i.e. non-increasing returns to scale[7].

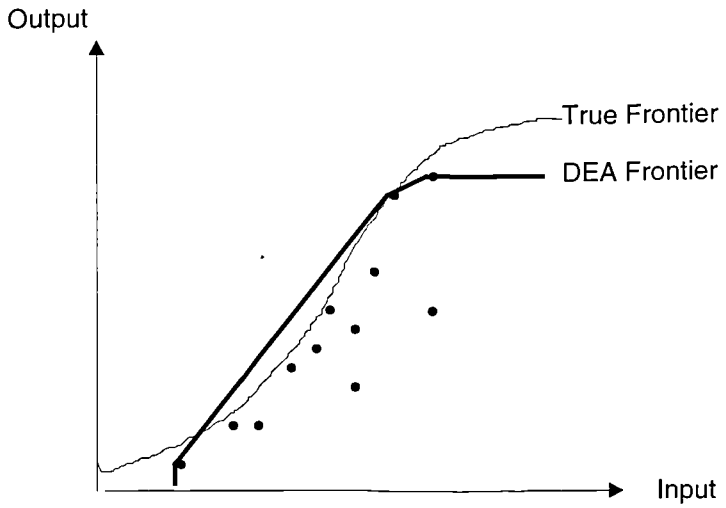
However, in DEA, the assumption of Shephard that the function must go through the origin is removed by allowing output to be gained only after a certain level of input is achieved. This allows the convexity assumption to be imposed while still allowing for increasing returns to scale.

In a similar way, the SF method can impose a certain functional form which gives negative values of output until a certain level of input is achieved. Below this level of input, the output is taken to be zero - negative output does not make sense. Therefore, the production frontier is actually made up of two distinct parts in both methods: a plane in the output = 0 space and the (positive portion of the) frontier estimated by the method. If the whole of this 'joint' frontier is considered it is actually non-convex. The convexity assumption in DEA only applies to the portion of the whole frontier which is being estimated using the observed DMUs. This 'joint' frontier has a similar shape to the S-shaped curve shown in Figure 2-8.

[7] There is no way that a function through the origin enclosing a convex set can exhibit increasing returns to scale.

Therefore, the only problem that may be encountered by the assumption of convexity in DEA is that some, or all, of the DMUs being assessed lie in the lower part of the S curve below point A in Figure 2-8: Then the assumption is violated by the data. The region of non-convexity will be given efficiency estimates by DEA which are less than the true values (See Figure 2-9).

Figure 2-9. The convexity assumption in DEA



In Figure 2-9 there are several units in the lower portion of the S-shaped curve. This is the non-convex region of the technology and most of the units in this region will be given estimates under DEA which are too low.

If the SF method does not impose convexity (e.g. by using a translog function as the estimating function) then this problem should not be encountered by the SF method.

In order to identify whether this problem may be occurring, the 'distance' between successive efficient DMUs needs to be examined. If at small-scale sizes[8] there is a region where the distances between successive efficient units are very large, the frontier between these units may not be well specified.

One way to remove the convexity assumption from DEA is by using the Free Disposal Hull (Deprins, Simar and Tulkens (1984)) method instead. However, rather than removing the general convexity assumption across the whole technology, the assumption is removed for individual facets of the frontier. This means that the general change from increasing to constant to decreasing returns to scale no longer holds.

[8] See Chapter 4.

Hypothesis 7 (not illustrated)

If the true technology is non-convex then the regions of non-convexity will be given estimates such that $E_{SF} \cong E_{TRUE} > E_{DEA}$ unless SF also imposes a convex estimating function, in which case the estimates will be such that $E_{TRUE} > E_{SF} \cong E_{DEA}$.

2.7.2 SF A4 does not hold: There is correlation between the inputs and the inefficiency term

In a regression analysis the exogenous variables are always assumed to be independent of the error. This may not be true in an efficiency analysis (see Gong and Sickles (1993)). If correlation does exist, the estimates will be biased in the SF methods but not in the DEA method (see Banker et al. (1996)). This possibility will not be considered in the analyses but the conclusions given in Chapter 8 will need to be modified if this may be a problem.

2.7.3 SF A5 does not hold: The inefficiency is in the inputs rather than the outputs

Banker et al (1988), Gong and Sickles (1992) and Banker et al (1993) assign inefficiency only to the dependent variable whereas Arnold et al. (1996) assigns it to the inputs.

Whenever a regression method is used, the error term is always assumed to be affecting *only* the endogenous variable. What happens

if the error is also in the exogenous variables? Arnold et al. (1996) and Cooper and Tone (1997) considered this problem and found that the SF results are very adversely affected. Once again, this possibility will not be considered here.

2.8 Conclusions

In this chapter, seven hypotheses have been given describing how each of the methods will be affected by the validity of the underlying assumptions. The next five chapters will consider some of these hypotheses and illustrate how the nature of the deviations between the two estimated methods can be used to draw conclusions about the nature of the true technology.

In order to test the hypotheses, simulated data sets will be used. This enables us to manipulate the properties of the underlying data set and gives the true efficiencies of the observed data. It is then straightforward to see which of the two methods is performing better for an individual unit.

It is possible for a single DMU to be given an estimate such that $E_{DEA} \cong E_{SF} \cong E_{TRUE}$ but this could be due to the effects of several non-valid assumptions cancelling each other. For $E_{DEA} \cong E_{SF} \cong E_{TRUE}$ across the whole technology all assumptions in each method must be valid. Hence, whenever testing the hypotheses we must begin with an

underlying data set that does not violate any of the assumptions and therefore gives very good estimates across the whole technology. Once it has been established that none of the assumptions has been violated, it is possible to manipulate the data set to test a specific hypothesis. In this way we will be sure that only one assumption is being tested.

The next chapter will consider the first two hypotheses where differences occur for units in any region of the technology. Chapters 4, 5 and 6 consider Hypotheses 3, 4, 5 and 6 where the variation of fit occurs in regions of different scale size. Chapter 7 investigates Hypothesis 6 where the variation of fit occurs across input mix and Chapter 8 summarises the results from these chapters.

Chapter 3

Investigating differences across the whole technology

3.1 Introduction

Now that the hypotheses for the effect of the properties of the data on the performances of the methods have been outlined, this chapter will investigate the cases where the effects occur across the whole technology.

The assumptions that can affect units in any region of the technology are those about the nature of the error term:

1. the assumption of no noise in the DEA method, and
2. the joint distribution of the random noise and the inefficiency distribution assumption in the SF method.

The next section will investigate the effects of random noise on the performance of the methods. It is shown that the noise affects DEA to a much greater extent than the SF method. However, the SF method is not unaffected by the presence of noise in the data.

In Section 3.3, the effects of incorrect assumptions about the nature of the inefficiency distribution in the SF methods are investigated. It is shown that for a half normal underlying inefficiency distribution, the assumed distributions of half normal, truncated normal or exponential all perform reasonably similarly and give good estimates when no other assumption of the SF method is violated. However, when the underlying inefficiency distribution is uniform, the performance of the SF

method is adversely affected when any one of the three assumed distributions is used. Obviously, the underlying distribution of the inefficiencies need not be either uniform or half-normal. However, it is shown that a distribution (i.e. uniform) that is very far from the assumed distribution in the method (i.e. half-normal, truncated normal, or exponential) may be identified by a comparison of the SF results with those of DEA.

In order to investigate these assumptions we will use data sets generated according to Data Generating Processes (DGPs) A, B and C from Appendix 2. Both DGP A and C have zero, low and high levels of random noise whereas DGP B has only low levels of noise. DGP B and DGP C both have an underlying half-normal inefficiency distribution. DGP A has two different underlying inefficiency distributions - half-normal and uniform.

Table 3-1. Summary of the data

	Inefficiency distribution	random noise levels
DGP A	uniform, half-normal	zero, low, high (additive)
DGP B	half-normal	low (multiplicative)
DGP C	half-normal	zero, low, high (additive)

Results from three DGPs have been included to ensure that the conclusions hold more generally than for a particular DGP. In the next section, the effect of the level of underlying random noise and its generating process (i.e. whether the random noise is additive or multiplicative) will be investigated. In Section 3.3, the effect of the underlying inefficiency distribution on the SF method will be investigated.

3.2 The effect of random noise on the performance of the methods

Random noise can occur due to any factors which are out of the control of the DMU, such as the weather, or it could be due to measurement errors or any misspecification in the model being used, e.g. omitted variables.

However, if random noise is present in the data, the two approaches handle it in very different ways. DEA ignores any random factors and attributes any deviation from the frontier to inefficiency. If we assume that the random noise term is symmetric, i.e. it is just as likely that the random noise would increase the true output as decrease it, then the DMUs on the true frontier will be pushed up and down in roughly equal numbers by the effect of the random noise. This will always lead to the estimated frontier lying above the true frontier whenever there is any random noise in the data and the data set is reasonably large. This in

turn leads to the efficiency estimates under DEA generally being less than or equal to the true values except for the few units which have very positive random noise.

In the case of the SF approach, the random noise is incorporated into the method by using a composed error. So, we would expect that the SF method would identify the random noise and the efficiency estimates would be good even for high levels of noise. However, the efficiency estimates are given by $E(u|\varepsilon)$ (see equation (1-20)) i.e. the efficiency estimated by the SF method is conditional on the total error and it is not apparent how this will affect the performance of the method.

Hypothesis 1

If the data contains random noise, then the whole technology will be given estimates such that $\overline{E}_{TRUE} \cong \overline{E}_{SF} > \overline{E}_{DEA}$.

When there is no random noise in the underlying data (very unlikely in practice) we would expect both SF and DEA to perform well when all other assumptions of the methods are valid. Once random noise is introduced we expect the DEA results to be less than the true efficiencies.

The effect of random noise on the results of the methods will be investigated by testing the methods with three different levels of noise which we will call zero, low and high levels.

3.2.1 Generating the random noise

Following Banker et al. (1988) we choose to introduce low random noise so that 95% of the observed outputs lie within ±10% of the actual outputs, and high random noise so that 95% of the observed outputs lie within ±40% of the actual outputs.

3.2.1.1 Multiplicative random noise

Consider the case of multiplicative error. Let the true output, y_{true} be given by $y_{true} = f(x; \beta)$. The efficient frontier is then given by $\tilde{y} = y_{true}e^{v}$ where $\tilde{y}$ is the efficient output and e^{v} is the multiplicative random noise term. Now, let $e^{v} \sim N(1,\sigma_v^2)$. This gives the mean of the efficient output to be the same as the mean of the true output and the standard error of the residuals to be $y_{true}\sigma_v$[1].

Suppose that there is no inefficiency in the data set. Then the observed output values would be the efficient outputs $\tilde{y}$. An ordinary

[1] Note that this gives heteroskedastic errors, i.e. the error increases as the output increases. Once the log-linear specification is used in the estimation method, the error becomes homoskedastic.

least squares regression could then be applied to this data set to obtain a regression line for the true output. For a single input, x, the 95% prediction interval for a specific true output, y_p, is given by

$$y_p = \tilde{y}_p \pm t_{0.025}\, y_p \sigma_v \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}} \qquad (3\text{-}1)$$

where n is the sample size of the data, x_p is the value of the input level chosen and $\tilde{y}_p$ is the efficient output including random noise, for input level x_p.

Note that for large values of n, the last two terms in the square root become very small and the square root term tends to the value 1. So for large samples

$$y_p \cong \tilde{y}_p \pm t_{0.025} y_p \sigma_v \qquad (3\text{-}2)$$

Therefore, in order to ensure that this interval is approximately ±10% of the true output (for low random noise), just set

$$t_{0.025} y_p \sigma_v \cong 0.1 y_p, \text{ or } \sigma_v \approx 0.05. \qquad (3\text{-}3)$$

(assuming that for large samples, $t_{0.025}$ is very approximately equal to 2).

Similarly for high random noise, $\sigma_v \cong 0.4/t_{0.025} \approx 0.2$.

This method has been used to generate the random noise term for DGP B.

3.2.1.2 Additive random noise

Now consider the additive case. Once again the true output, y_{true}, is given by $y_{true} = f(x;\beta)$ but the frontier now is $\tilde{y} = y_{true} + v$ where $\tilde{y}$ is the efficient output and v is the random noise term. Now, let $v \sim N(0,\sigma_v^2)$. This gives the mean of the efficient output to be the same as the mean of the true output and the standard error of the residuals to be σ_v.

Now, for large n, the 95% prediction interval for a particular true output, y_p is approximately

$$y_p \cong \tilde{y}_p \pm t_{0.025}\sigma_v \qquad (3\text{-}4)$$

In this case, to ensure that 95% of the true outputs lie within ±10% of the observed outputs, we need

$$t_{0.025}\sigma_v \cong 0.1y_p \qquad (3\text{-}5)$$

which gives

$$\sigma_v = 0.1y_p / t_{0.025} \qquad (3\text{-}6)$$

This method has been used to generate the random noise for DGPs A and C, setting y_p to be the mean value of the true outputs. This error is homoskedastic at this raw data level. However, in the log form the error becomes heteroskedastic.[2]

By comparing the effect of the random noise on the SF results of DGPs A and C with those of DGP B (all having half-normal underlying inefficiency distributions), we can see whether the assumption of an homoskedastic error in the log form has an effect on the estimates.

Note that in order for the prediction interval to always include ±10% of the true output, the random noise term must be heteroskedastic. For the multiplicative random noise term this is fine as long as a log form is

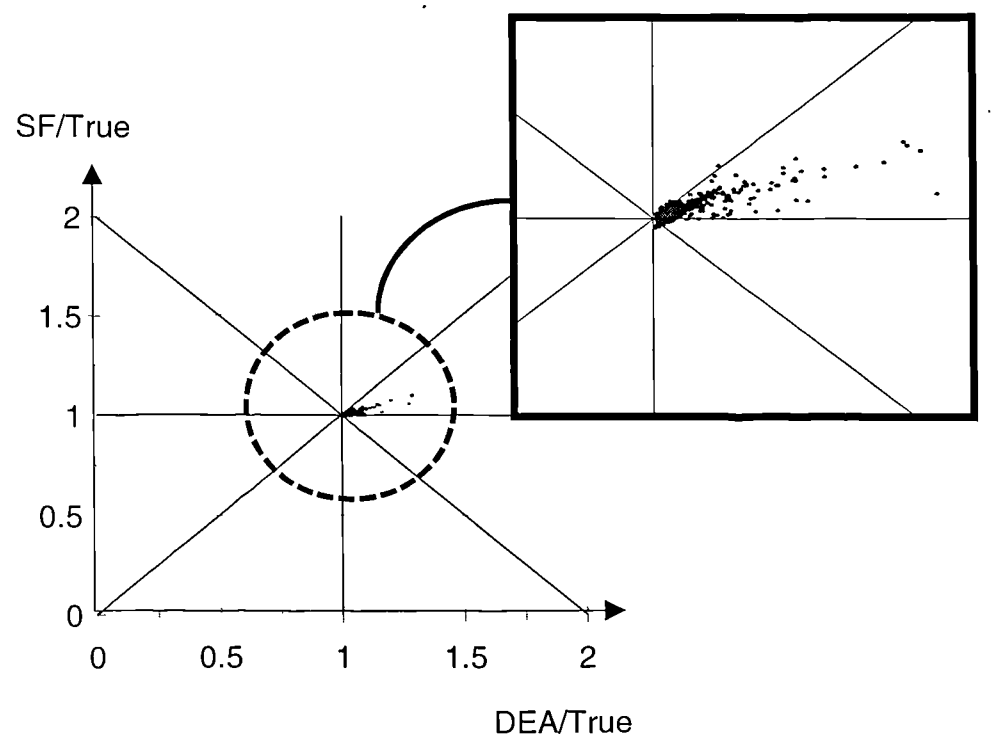
---

[2] For an additive error of the form $y = f(x;\beta) + v$, the log form will become $\ln y = \ln(f(x;\beta) + v)$. Taking a first order Taylor expansion of this expression about $\ln(f(x:\beta))$ gives $\ln y = \ln(f(x;\beta)) + \frac{v}{f(x;\beta)} + O(2)$.

used for estimation. For the additive case, we have chosen the true mean output to set the level of the variance of the noise. This will lead to a heteroskedastic error in the log form as explained above. There does not appear to be a good reason for *a priori* assuming one of these scenarios to be more likely than the other. For this reason, both methods have been used to generate the random noise in order to see how much effect the assumption has on the results.

3.2.2 Results - The effect of random noise on the methods

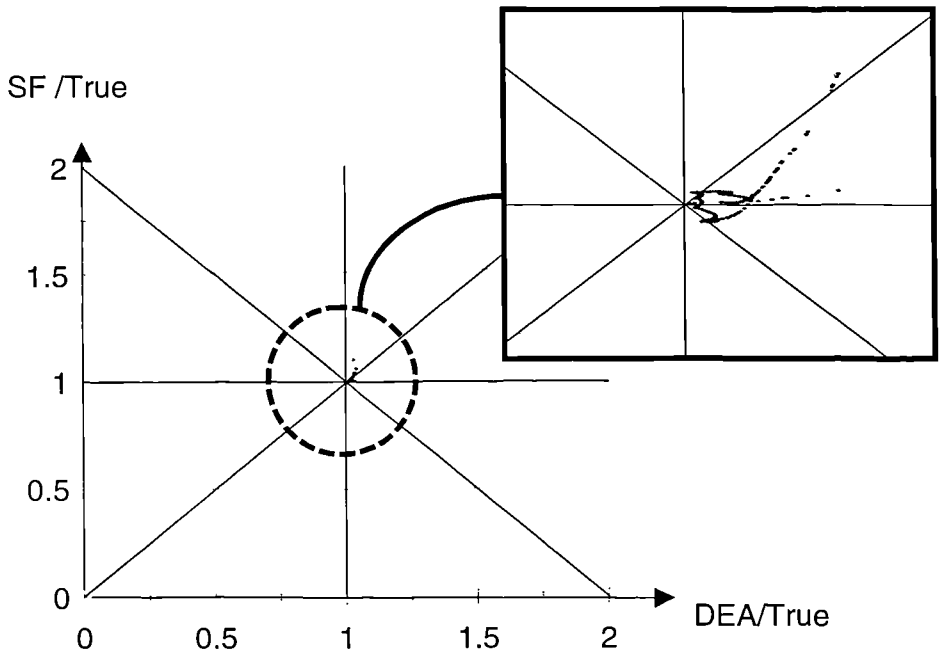
Figure 3-1. DGP A (no random noise and truncated-normal[3] inefficiency assumption)



[3] In all cases in this section, the SF method has assumed a truncated normal distribution for the inefficiency distribution as this assumption gave a full set of results.

Before introducing any random noise into the data, first notice in Figure 3-1 and Figure 3-2 that both methods perform very well for data sets which have no random noise.

Figure 3-2. DGP C (no random noise and truncated-normal inefficiency assumption)



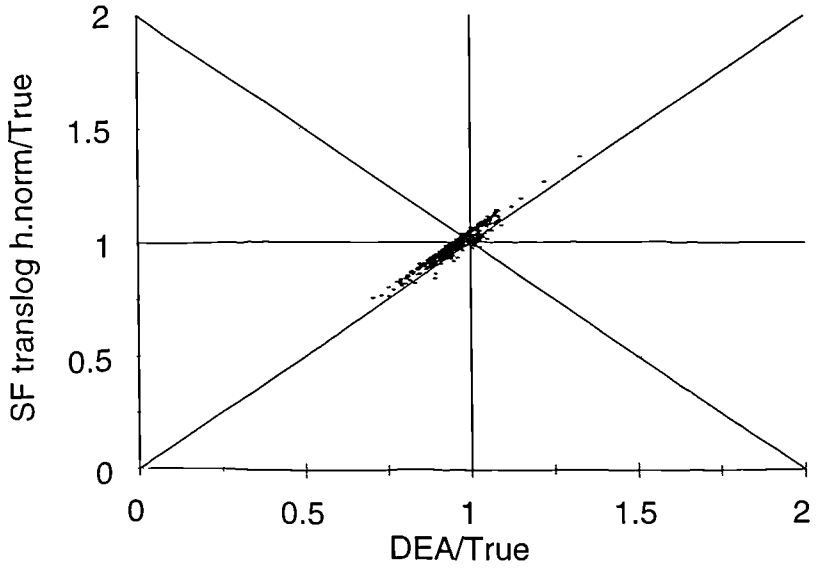
For each of these graphs, the ratio of the estimated SF efficiency (using a translog function in order to avoid functional misspecification) to the true efficiency, is plotted against the ratio of the estimated DEA efficiency to the true efficiency for each DMU in the data set.

DGP C gives slightly better results in this case as all the assumptions of the methods are met, whereas DGP A involves a region of non-

convexity, so the DEA results are not as good. Note that in all cases, when all assumptions of DEA are met, there will still be some finite sample error in the estimates leading to the DEA estimates being greater than or equal to the true efficiency values. This is due to the fact that the DEA frontier is a piecewise linear approximation to the true frontier. The estimates have been shown to be asymptotically consistent (i.e. as the sample size increases, the sample error decreases (Banker (1993)).

Similar graphs can be plotted when the low and high random noise levels are introduced - Figure 3-3 and Figure 3-4.

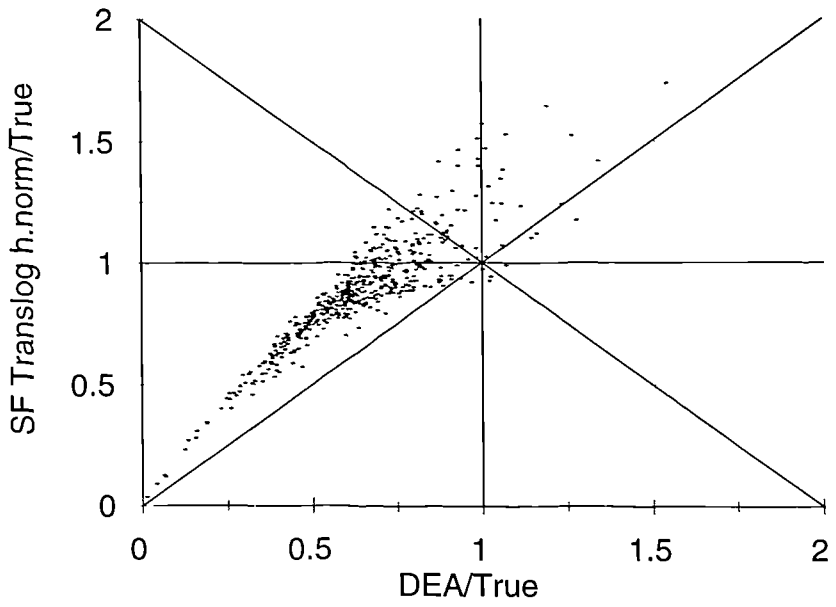
Figure 3-3. DGP C: The effect of low random noise on the results



The closer the points are to the 45° line through the origin, the less there is to choose between the DEA and SF estimates.

For low random noise, we find that the DEA and SF results are similar, and there are a reasonable number of units for which DEA outperforms the SF method and vice versa. (The graphs for DGPs B and C under low random noise are very similar to those shown here.) Note that the results are found on both sides of the vertical line through (1,0), so for low levels of random noise the DEA results are not *all* less than the true efficiency values. This is true of all the data sets for low levels of noise. This is because, although the frontier is shifted up, some units are given efficient outputs that are much greater than their true outputs due

Figure 3-4. DGP C: The effect of high random noise on the results



to the random noise. When this difference is greater than the shift of the frontier, the efficiencies will be overestimated by DEA. The number of these units decreases as the level of noise rises.

Under high random noise, the units are biased away from the 45° line, towards the horizontal line through (0,1). This is the line for which the SF estimates are equal to the true values. Now the majority of the units are in a region where the SF estimates are closer to the true values than the DEA estimates (see Figure 2-4(d)). Almost all the results are now to the left of the vertical line through (1,0) showing that the majority of units are given estimates such that $E_{TRUE} > E_{DEA}$.

The values of the mean absolute deviations (MAD) of the estimated efficiency values from the true efficiency values are given in Table 3-2 when low and high random errors are introduced.

Table 3-2. Mean absolute deviations of the estimates from the true efficiencies

Estimating method	Level of noise	DGP A	DGP B	DGP C
DEA	Zero	0.01669	-	0.00700
	Low	0.05325	0.05145	0.05135
	High	0.22826	-	0.29175
SF (translog, truncated normal)	Zero	0.00659	-	0.00287
	Low	0.03996	0.05253	0.04150
	High	0.14219	-	0.14609

Clearly, the random noise has a much greater effect on the DEA results than the SF results. For zero and low levels of noise both SF and DEA perform similarly, although in each case the SF estimates are slightly better on average than the DEA estimates (except for DGP B which has been generated with a multiplicative random noise term). However, for high levels of noise, DEA is affected to a much greater extent than SF.

The DEA method is not allowing for any random noise. The SF method does allow for random noise but it is not possible to completely separate the noise from the inefficiency. The efficiency estimates under DEA are much less than the true values when there is high random noise. The SF values are much closer to the true values than

Table 3-3. Correlation coefficients: DGP C, half-normal underlying inefficiency

	Level of noise	DEA	Translog SF	True
DEA	zero	1		
	low	1		
	high	1		
Translog SF	zero	0.99538	1	
	low	0.98024	1	
	high	0.92636	1	
True	zero	0.99023	0.99849	1
	low	0.92584	0.94615	1
	high	0.55966	0.60309	1

DEA. However, it is clear from Table 3-3 that for high levels of noise the SF efficiency values have similarly poor levels of correlation to the true values as DEA, while the estimates are highly correlated with each other.

For example, under high random noise, the level of correlation between the DEA and true values is 0.56, while the correlation between SF and the true values is 0.60. The correlation between the DEA and SF estimates under high random noise is 0.93, much higher.

So the hypothesis that $\overline{E}_{TRUE} > \overline{E}_{DEA}$ when high random noise is present in the data has been illustrated. However, we have also shown that as the level of noise increases, the SF method will begin to attribute more of the total error to be inefficiency rather than random noise. Therefore the estimates will be such that $\overline{E}_{TRUE} > \overline{E}_{SF} > \overline{E}_{DEA}$ when there is high noise in the data and all other assumptions of the methods are met.

The results for DGP B were included in Table 3-2 to illustrate that the results of the SF method do not seem to be affected to a large extent, by whether the random noise is correlated to the output, i.e. heteroskedastic. DGP B has a multiplicative random noise term at the raw data level. Therefore, as shown earlier, at the logged data level the error term becomes homoskedastic. DGPs A and C on the other hand,

have additive error terms at the raw data level, which gives heteroskedastic errors at the logged data level. The results in Table 3-2 for each of the DGPs A, B and C give similar results for the SF method in comparison with the DEA results. So, the assumption of an homoskedastic error in the SF method does not seem to affect the performance of the method to any great extent.

3.3 The Inefficiency distribution

Before the SF method can be applied, an assumption must be made about the distribution that the inefficiencies take. We would like to know how this assumption impacts on the estimated inefficiencies.

As noted by Coelli (1994), "It appears that the vast majority of applied papers involve the estimation of a single equation half-normal stochastic frontier. [Bauer (1990,p53) and Bravo-Ureta, Pinheiro (1993, p97) have also made this observation.]" It seems likely that this is the most popular method due to the fact that this was the first method proposed by Aigner, Lovell and Schmidt (1977). For a half-normal inefficiency distribution, the probability density of DMUs decreases monotonically as the inefficiency increases, which may be unlikely in practice. Also, as the truncated normal includes the half normal as a special case, the truncated normal should be able to identify an underlying half-normal inefficiency distribution.

The three assumptions we will consider are a half-normal distribution, a truncated-normal distribution and an exponential distribution (see Appendix 1 for references).

In Chapter 2 the following hypothesis was formulated:

Hypothesis 2

If the assumed inefficiency distribution is not the same as the true inefficiency distribution, the SF method will compensate by allocating more of the total error to random noise. The differences between the true and estimated efficiencies will vary as the true inefficiency changes.

Once again, DGPs A, B and C will be used to investigate the effect of the inefficiency assumption.

3.3.1 Results - The effect of different inefficiency assumptions on the SF method

In Table 3-4, the results obtained by assuming that the inefficiencies are distributed with a half-normal distribution are compared with the results obtained with a truncated-normal or exponential distributional assumption for data sets from each of the DGPs.

The SF method clearly gives much worse results when the underlying inefficiency distribution is uniform. The method, using one of the three

assumptions in the table, seems to rely heavily on the underlying inefficiency distribution being biased towards zero.

Table 3-4. Mean Absolute Deviations. All estimated using a translog SF function and no random noise

Inefficiency assumption	True inefficiency Uniform DGP A	half-normal DGP A	half-normal DGP C
half-normal	**	**	0.00287
truncated normal	0.13558	0.00659	0.00370
exponential	0.09808	**	0.00376

** MLE gave type II errors.

When the underlying inefficiency distribution is half-normal, the assumption of the inefficiency term has very little effect on the results as shown by the results for DGP C in Table 3-4. A much more in depth study would be needed to say whether this is true in all cases. Note that the half-normal assumption is slightly better than the other assumptions as would be expected.

Figure 3-5 below shows the ratios of the DEA estimates to the true efficiency values in comparison with the ratios of the SF estimates to the true efficiency values. Clearly, the SF estimates have been affected by the uniform distribution of the true inefficiencies when the method has assumed a truncated-normal inefficiency distribution. (Compare Figure 3-5 with Figure 3-1 where the underlying inefficiency distribution was half-normal.) The DEA results do not appear to have

been affected. The SF estimates are now all much less than the true values.

Figure 3-5. DGP A: underlying uniform inefficiency (no random noise and truncated-normal inefficiency assumption)

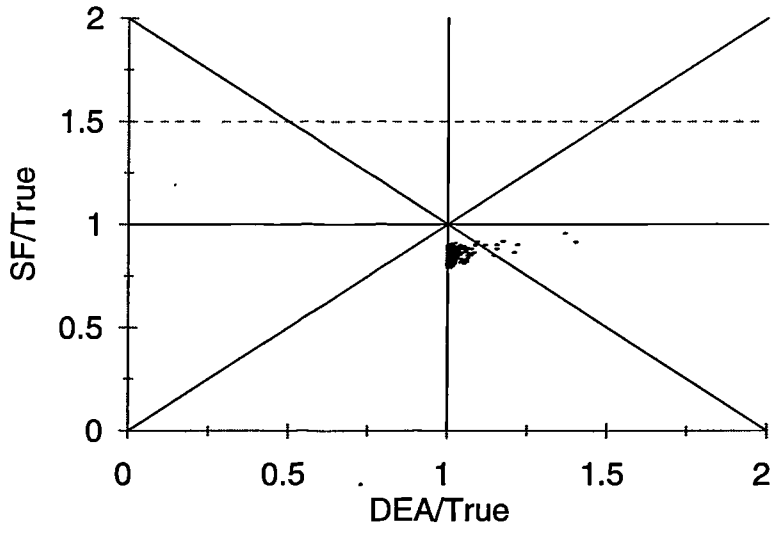
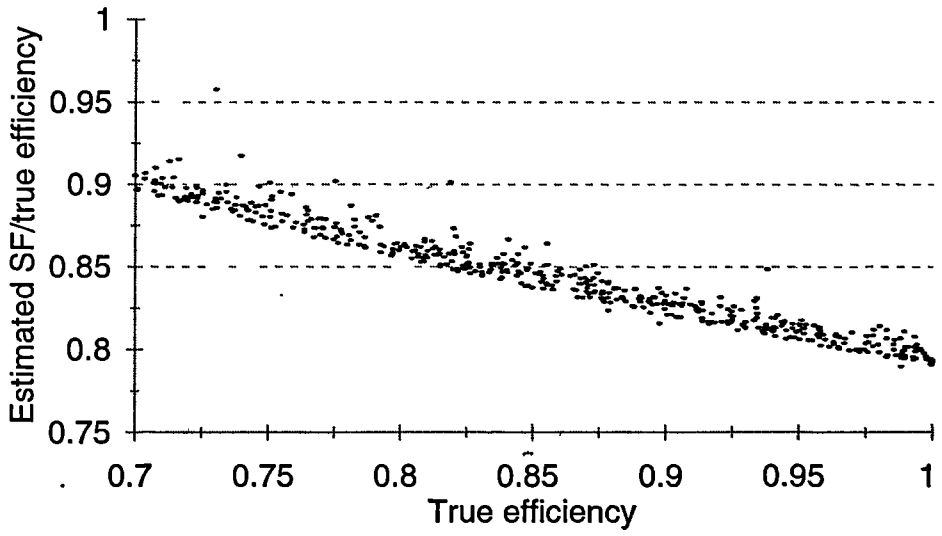


Figure 3-6. The effect of an underlying uniform inefficiency distribution (DGP A)

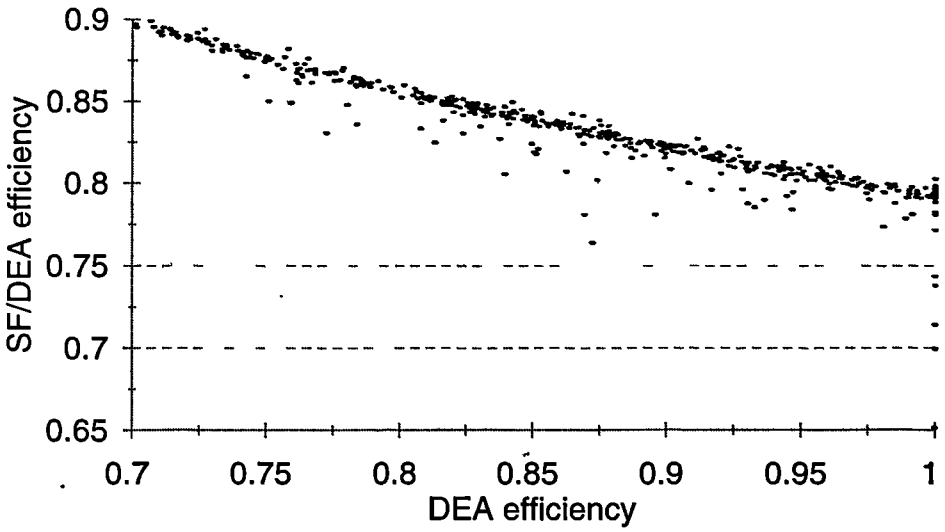


In Figure 3-6, the ratio of the estimated SF efficiency is plotted against the true efficiency for the data set from DGP A with an underlying uniform inefficiency distribution.

From this graph, there is clearly a relationship between the deviation of the estimated efficiency from the true efficiency and the level of true efficiency. The translog SF method gives closer estimates for the most inefficient units than the efficient units.

In order to identify possible misspecification of the inefficiency term in the SF method when the true efficiencies are unknown (i.e. in a real data set) the SF and DEA efficiency estimates can be compared. This is done in Figure 3-7. It is clear from this graph that the differences

Figure 3-7. Uniform underlying inefficiency - a comparison between the SF and DEA estimates (DGP A)



between the estimates are varying as the level of the DEA estimated efficiency changes. The SF and DEA efficiencies are clearly giving the closest estimates for the least efficient units. As there is no reason for the DEA method to be affected by the level of efficiency, it is possible to use such a graph to indicate whether there is any misspecification in the assumption of the inefficiency term in the SF method. (Similar results are found which ever of the three inefficiency assumptions are made in the SF method when the underlying inefficiency is uniform.)

3.4 Conclusions

In this Chapter the effect of the assumptions about the error terms in each method on the performance of the method has been investigated. The DEA method is affected by any random noise in the data. For all levels of random noise the majority of DMUs are given lower estimates under DEA than the true values, although for low levels of noise there are a significant proportion of units for which DEA overestimates the efficiency.

The usefulness of these results lies in knowing that in a real data set, if it is found that almost all the DEA estimates are less than the SF estimates and there does not seem to be any pattern to the differences, it is possible to conclude that there is likely to be random noise in the data.

It is possible to have some idea of the amount of noise associated with the data by considering what type of data it is: Banker, Gadh, and Gorr (1993); "Both low and high measurement errors are likely encountered regularly in applications. For example, a manufacturing firm's output can be easily counted and inputs are directly controlled by the firm and so are easy to identify. Low errors are expected in such cases. In contrast, public sector agencies and not-for-profit organisations produce services for which it is difficult to quantify outputs and identify inputs." It seems to have been assumed in previous studies on actual data that when the two methods give similar results they must both be performing well. It has been implied that, because the two methods have such different underlying structures that when they agree, it points to them both doing well. It is assumed that SF methods will take into account any random noise in the data and therefore, when there is some noise, SF methods will perform much better than DEA. Our results show that although the SF estimates are closer in value, on average, to the true estimates for high random noise, the correlation is as poor as that between DEA and the true values.

The effects of misspecification in both of the random noise components in the SF method have been investigated individually. It has been shown that heteroskedasticity in the random noise does not appear to affect the results to any great extent. The choice of the inefficiency distribution between half normal, truncated normal and exponential also

has little affect on the results. When the underlying inefficiency has a half-normal distribution, a half-normal, truncated-normal or exponential assumption in the SF method produces good results. When the underlying inefficiency distribution is uniform, all three of these assumptions produce poor results. It has been shown that the ability of the SF method to identify the true distribution can be investigated by comparing the SF and DEA estimates across the DEA estimates.

This chapter has investigated the effect of assumptions that affect the performances of the methods across the whole technology. The general conclusions are:

- If all assumptions are met in the DEA method (i.e. there is no random noise in the data), the DEA efficiency estimates will all be equal to (or slightly greater than) the true values.
- The only possibility that has been identified for the DEA estimates to be less than the true values on average across the whole technology is when there is random noise in the data. This also leads to the SF estimates being less than the true values but greater than the DEA estimates.

Random noise in the data $\Rightarrow \overline{E}_{TRUE} > \overline{E}_{SF} > \overline{E}_{DEA}$

The only exception to this is if there are low levels of noise in the data and the SF method correctly identifies the magnitude of this noise on average

Low random noise in the data $\Rightarrow$ either $\overline{E}_{TRUE} > \overline{E}_{SF} > \overline{E}_{DEA}$

or $\overline{E}_{TRUE} \cong \overline{E}_{SF} > \overline{E}_{DEA}$

- Another possibility for the SF estimates to be less than the true efficiency values across the whole technology is if the inefficiency distribution is incorrectly specified in the SF method. In this case the DEA estimates are not affected, as DEA makes no assumptions about the distribution of inefficiency. This gives

Incorrect inefficiency assumption in SF $\Rightarrow$

$E_{DEA} \cong E_{TRUE} > E_{SF}$ across the whole technology

In the next three chapters, differences between the estimates across scale size and in Chapter 7, differences between the methods across input mix, will be investigated.

Chapter 4

Scale size for the single-output,

multiple-input case

4.1 Introduction

In order to investigate differences between the methods across scale size, it is necessary to have some idea of what is meant by scale size, particularly when multiple variables are involved.

What is a small-scale unit and what differentiates it from a large-scale unit? Is there a unique definition of size that can be applied to any data set? Does it depend on the orientation that is chosen to measure efficiency? Forsund (1996);

> "The empirical interest in [scale] centres around whether there are economies or diseconomies of scale, the implications for market structure and conduct, and government regulation policy concerning price, etc. Should the size of the units under investigation be expanded or contracted as policy conclusions? What is the optimal size of a hospital, bank, industrial firm, etc.? What do we mean by size in a multiple-output multiple-input setting?"

It is this final question which will be addressed here for the single output case, while the general case for multiple outputs will be discussed in Chapter 6. In this chapter we will define a **cross-mix scale size** for the single output case and give an example to show how it is calculated. The next chapter will then give an example to show how this definition of scale size can be used to learn more about the underlying data.

4.2 Defining scale size

If one DMU uses smaller amounts of all inputs and produces less output than a second DMU then it would seem reasonable to say that the first DMU is operating at a smaller scale size than the second. However, this definition of size becomes more complicated when only some of the variables are smaller, or when some are smaller and some are larger. What is necessary for one DMU to be considered to be operating at a smaller scale than another?

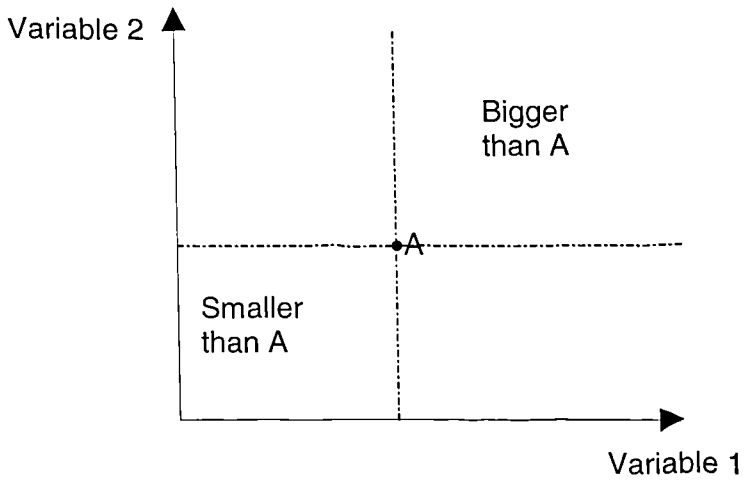
Before investigating the problems of defining scale size in production theory, first consider the general problem of defining what is meant by size. If we have a single measurable variable, e.g. length, it is straight forward to say whether one unit is larger than another unit: There are only three possibilities; the first unit has a smaller length than the second unit and we say that the first unit is smaller than the second unit; the first unit has the same length as the second unit and we say that the two units are the same size; or the first unit has a larger length than the second unit and we say that the first unit is larger than the second unit.

As soon as another variable is introduced, we encounter problems in defining size.

> "There is no way mathematically to well order more than two variables at the same time. ... One way around this problem is to combine all the measurements into a single 'figure of merit'. ... It's a way of adding up all the important attributes of something to arrive at a single number that can then be compared with other similar things." K. C. Cole (1998).

For example, suppose we want to compare the sizes of two pieces of paper. We are given the length and width of each piece. If the length and width of B are twice those of A, we can say that B is bigger than A. This is true of any units on the same ray through the origin in n-dimensional space. One unit, which has greater values for all the variables than a second unit, can be said to be a larger unit. This leads to our first proposition:

Figure 4-1. Comparing units with 2 variables relating to size



Proposition 1: If for two production units A and B, the values of the variables of B are all greater than the values of the variables of A, we can say that the scale size of B is greater than the scale size of A.

However, even with only two variables, this still does not give us a unique definition of size. If the width of one piece of paper is smaller than the other, but its length is larger, can we compare the sizes of the two pieces of paper? The only way that the sizes of any two units (in this case pieces of paper) can be compared is if we have some way of aggregating the variables.

In this example, the variables of length and width can be aggregated by multiplying the length and width of the paper to find its area. Then each piece can be allocated a size according to its area.

$$\text{Area} = \text{length} \times \text{width} = x_1 x_2 \qquad (4\text{-}1)$$

Note that this is a Cobb-Douglas function with increasing returns to scale. If we use the area to measure the size, then by doubling the length and the width, we quadruple the size. Another measure of size is

$$\text{Size of piece of paper} = \sqrt{\text{Area}} = \sqrt{\text{length}} \times \sqrt{\text{width}} = x_1^{1/2} x_2^{1/2} \qquad (4\text{-}2)$$

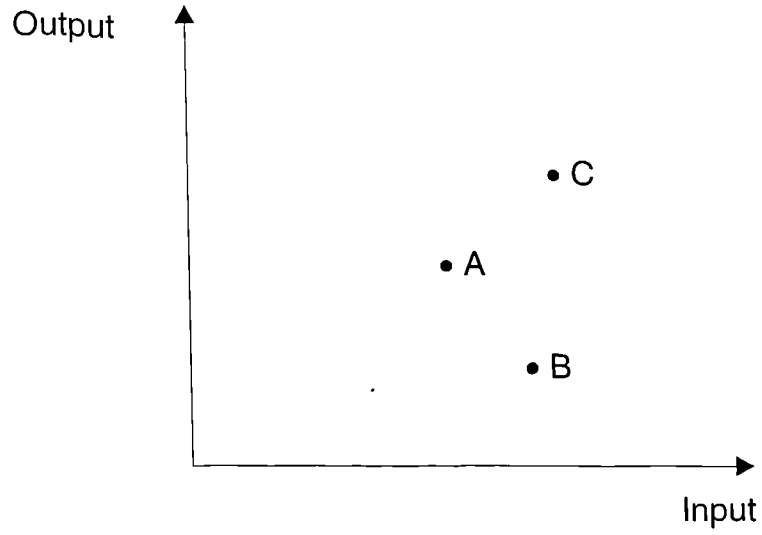
By using this method of aggregation, we find that by doubling both the length and the width of a piece of paper, we have doubled the size. This is always true if we use a CRS function for aggregating variables. In production theory, we will show that the CRS production function can be used for aggregating the variables into a single scalar measure representing the size of operation.[1]

In the single-output, multiple-input technology, the observed output is often used as a measure of the scale size of the DMUs. Note that this is the same as using Area to measure the size of pieces of paper as in (4-1). However, in the example where we were measuring the size of pieces of paper, all the units (pieces of paper) lay on the function given by equation (4-1). This is a precise relationship. Once we move into production theory, we encounter inefficiencies. In a production situation where the inputs are exogenous variables and the output is controllable (e.g. the output of turnover of a commercial outlet is controllable while the size of the market in which it operates (an input) is exogenous) the level of output may be reflecting inefficiency rather than scale size: A DMU producing a low level of output may be a 'large scale' DMU which is very inefficient. In this case, the observed output cannot be used to measure scale size as it is contaminated by inefficiency.

[1] Note that the term scale size is used to denote the scale of operation of a production unit. This should not be confused with the physical size of a unit, which may be one of the input variables in an efficiency analysis.

For the single-input, single-output production case shown in Figure 4-2, the scale size of unit C is larger than the scale size of unit A from Proposition 1. We can also say that the scale size of unit C must be greater that of unit B as both the input and output of C are larger than

Figure 4-2. The single-input, single-output case

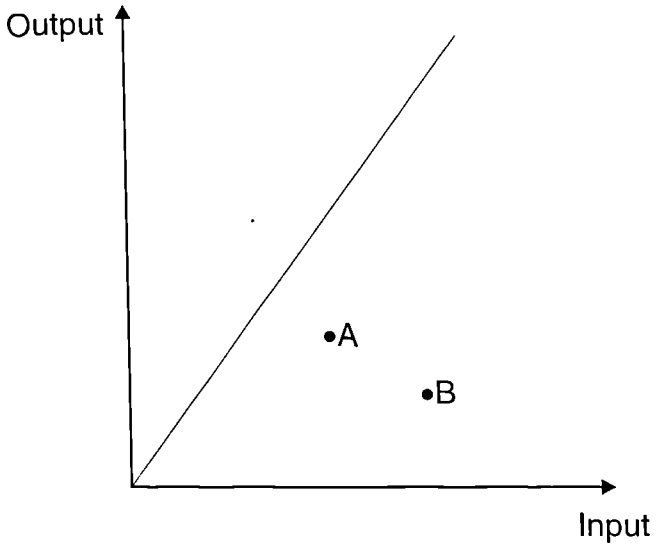


those of B. How can we compare the scale size of unit A and the scale size of unit B? In order to answer this question a definition of scale size across different input/output mixes is required.

It is only possible to compare the scale size of units across different rays in input-output space once they have all been projected onto the same ray. The way we choose to do this is to project onto the CRS frontier in the normal way, i.e. to eliminate the overall inefficiency in either an input or output direction. (If the VRS frontier is used, then the measure of relative scale size which we will later develop will be

dependent upon the chosen reference unit. By using the CRS frontier the relative sizes will be independent of the choice of reference. This means that the relative scale size developed here will be a transitive measure, i.e. if DMU A is found to be α times as big as DMU B, and DMU B is β times as big as DMU C, then DMU A is $\beta\alpha$ times as big as DMU C.)

Figure 4-3. The CRS frontier

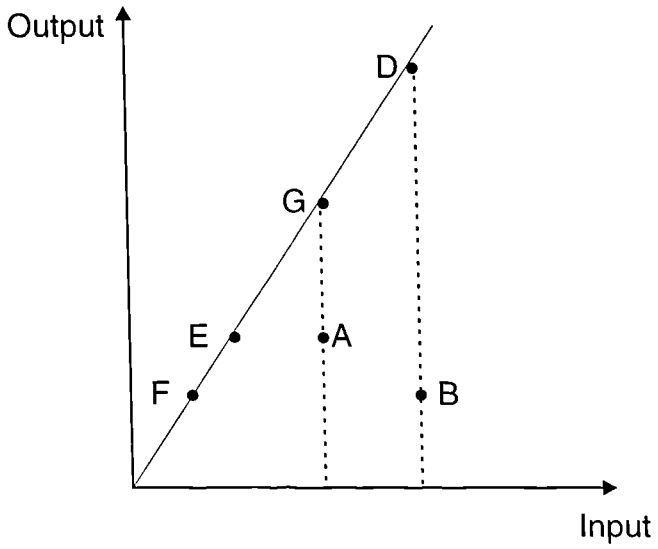


Note that, in Figure 4-3, if the input level is chosen to reflect the scale size of the units, then DMU A will be operating at a smaller scale than DMU B. However if the output level is chosen to reflect the scale size, then DMU A will be operating at a larger scale than DMU B. In a case where output is controllable but inputs are exogenous the low output of DMU B reflects output inefficiency. Conversely, in an input orientation, the low output level reflects the small-scale size of DMU B. This is

similar to the concept of scale inefficiency only making sense once the DMU has been projected onto the frontier. In this single-input, single-output example, once the orientation used to project the DMU onto the frontier has been decided, the definition of scale size is straightforward.

In the example above, if we choose to use an output orientation, the units will be projected to the points G and D (see Figure 4-4). Once the units have been projected it is clear that D is operating at a larger scale than G as it involves an increase in both inputs and outputs. Similarly, if the units are projected under an input orientation, they will be projected to points E and F. F is clearly operating at a smaller scale size than E.

Figure 4-4. The choice of orientation



Note that, once the units have been projected onto the CRS frontier, either the inputs or the outputs can be used to measure the relative scale sizes. For any two efficient units, A and B, on the same ray, it is clear that if the inputs of unit A are α times those of unit B then the outputs of unit A will also be α times those of unit B.

Now, if the unit in question has been projected onto the frontier in an output orientation then the efficient level of input will clearly be the same as the observed level of input. Therefore, once we have decided on an output orientation (because, for example, the inputs are exogenous) the *observed* input values can be used to define the relative scale sizes of all the units.

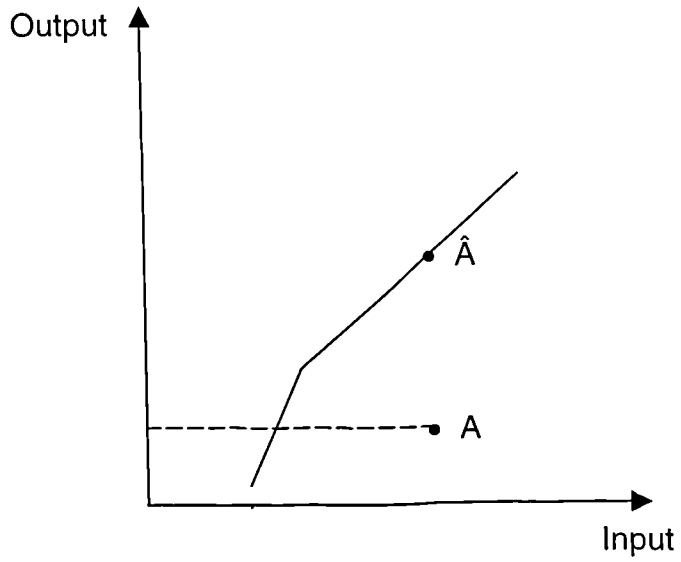
Note that the reverse is true for an input orientation, i.e. the scale size can be measured on the observed outputs but **not** the observed inputs.

In an output orientation, the *observed input values* can be used to define the relative scale sizes.

In an input orientation, the *observed output values* can be used to define the relative scale sizes.

The graph in Figure 4-5 illustrates why, when an input orientation is used to measure efficiency, scale size should be measured on outputs.

Figure 4-5. Scale efficiency and scale size



DMU A is scale inefficient on an input oriented measure of its efficiency under a VRS model. This means that the scale size of unit A is not the mpss. Lowering or raising its input level will only affect its VRS technical efficiency, its scale efficiency being constant. Only by raising its output can the DMU become scale efficient, at Â for example. In other words, the DMU has changed the amount of output it is using to eliminate the input scale inefficiency and reach the most productive scale size. This highlights why, for an input orientation, scale size should be measured on output. Similar arguments can be put forth as

to why, in cases where output is controllable and inputs are exogenous, scale size should be measured on input.

4.3 Using the Malmquist index to measure scale size

In this section it will be shown that the Malmquist quantity index[2] can be used to measure the relative scale sizes of a set of production units.

The Malmquist input quantity index (Caves, Christensen, and Diewert (1982)) is defined as the ratio of two distance functions:

$$M = \frac{D_I(y_1, x_2)}{D_I(y_1, x_1)} \qquad (4\text{-}3)$$

where the input distance function is defined for any input, output pair (x,y) as

$$D_I(y,x) = \max \{\lambda : (\frac{1}{\lambda}x, y) \in PPS\} \qquad (4\text{-}4)$$

Note that the input, output pair need not be in the PPS. In (4-3), the numerator of the index involves an unobserved input, output pair (y_1,x_2). This means, create a new unit which has the input levels of unit 2 and

[2] Note that this is not the same as the Malmquist *productivity* index which relates to production technologies in two different time periods.

the output levels of unit 1 and then measure the distance function for this unit against the original frontier (note that when measuring scale size, this will always be the CRS frontier).

The input distance function is obviously the same as the inverse of the input technical efficiency when the output in the distance function is equal to the output of the unit. The distance function will be equal to 1 for a unit which is efficient, greater than 1 for a unit which is inefficient and less than 1 for a unit which is infeasible (this is possible if the output level in the distance function is not the same as the output level of the unit).

Similarly, the output distance function is defined as

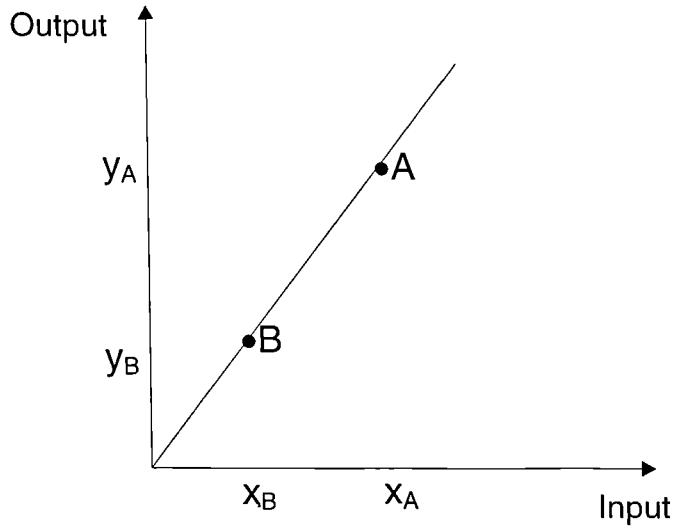
$$D_O(x,y) = \min \{\theta : (x, \frac{1}{\theta} y) \in PPS\}. \qquad (4\text{-}5)$$

$D_I^{CRS}(y,x)$ and $D_O^{CRS}(x,y)$ will be used to denote the distance functions relative to the CRS frontiers. (See Fare and Primont (1995) for a more in depth discussion of distance functions.)

4.3.1 The single input case

Consider the two DMUs A and B, for example, in Figure 4-6.

Figure 4-6. Scale size for CRS efficient units



Either the inputs or outputs can be used to measure the relative scale size as both A and B are CRS efficient, i.e. they already lie on the same CRS ray.

Let S(A) denote the size of unit A.

At A, the input of B, x_B, has increased by a factor x_A/x_B and the output of B has increased by a factor $y_A/y_B = x_A/x_B$.

So the scale size of unit A relative to unit B can be defined by

$$\frac{S(A)}{S(B)} = \frac{x_A}{x_B} = D_I^{CRS}(y_B, x_A) \qquad (4\text{-}6).$$
$$= \frac{y_A}{y_B} = D_O^{CRS}(x_B, y_A)$$

I.e. the scale size of DMU A relative to that of DMU B is given by the inverse of the DEA input efficiency of DMU A with output levels of DMU B or the inverse of the output efficiency of DMU A with input levels of DMU B.

The relative scale size measures the factor that the variables of B should be increased by to reach the levels of DMU A.

This leads to Proposition 2:

Proposition 2: If DMU A has CRS efficient output (or input) levels which are all greater than those of DMU B then DMU A must have a larger scale size than DMU B.

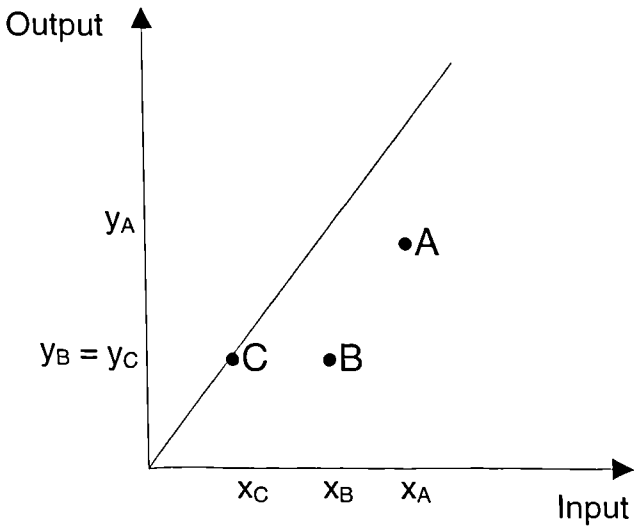
What happens if either A or B is inefficient? The orientation now becomes important. In Section 4.2 it was argued that in an output orientation, the scale size should be measured on the inputs. In this case, if B is efficient but A is inefficient, then the result in (4-6) for the ratio of inputs will obviously not change. That is, if A is inefficient then

the projection in the output direction to make A efficient will only change the output levels of A: The input distance function, $D_I^{CRS}(y_B, x_A)$, will remain the same. Vice versa for the output distance function. Therefore, the result for the input distance function in (4-6) holds if A is inefficient in an output orientation and the result for the output distance function holds if A is inefficient in an input orientation.

From now on, take the input to be exogenous, i.e. we are working in an output orientation.

The situation becomes somewhat more complicated if DMU B is inefficient. (See Figure 4-7 for example.)

Figure 4-7. Scale size for inefficient units



If B is inefficient then the scale size of A relative to C (efficient unit with the same output as B) is given by

$$\frac{S(A)}{S(C)} = \frac{x_A}{x_C} = D_I^{CRS}(y_C, x_A) = D_I^{CRS}(y_B, x_A) \qquad (4\text{-}7)$$

and the scale size of B relative to C is given by

$$\frac{S(B)}{S(C)} = \frac{x_B}{x_C} = D_I^{CRS}(y_C, x_B) = D_I^{CRS}(y_B, x_B). \qquad (4\text{-}8)$$

Therefore the scale size of unit A relative to B is given by

$$\frac{S(A)}{S(B)} = \frac{S(A)}{S(C)}\frac{S(C)}{S(B)} = \frac{D_I^{CRS}(y_B, x_A)}{D_I^{CRS}(y_B, x_B)}. \qquad (4\text{-}9)$$

This is the Malmquist input quantity index. Therefore, the Malmquist input quantity index can be used to measure cross-mix scale size whenever we have exogenous inputs. This index was discussed in Caves, Christensen and Diewert (1982) for comparing the input vectors of two different DMUs, but throughout, each unit was assumed to be operating at the frontier. Here we will show that this index can be used to compare the input vectors of any two units to define relative scale sizes.

Similarly, in an input efficiency orientation the Malmquist output quantity index can be used to measure scale size.

The choice of a CRS technology to measure the scale size does not depend on the underlying technology. Even if the underlying technology is VRS, the CRS technology must be used to measure the scale size in the same way that the CRS frontier must be used when measuring the Malmquist productivity index (Maniadakis and Read (1997)). If the VRS technology is used, the scale size will be compounded with scale inefficiency effects. Any CRS technology can be used in the single-input, single-output case as the frontier for the distance function. The relative sizes will obviously be unaffected.

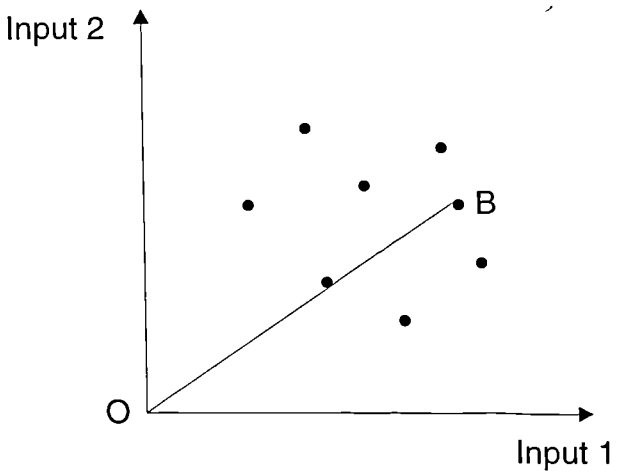
Once it has been decided that scale size is to be measured on input or output variables, the question arises as to how scale size is to be measured when multiple variables are involved.

4.3.2 The multiple-input case

Consider, for example, the input space of a two-input, one-output technology shown in Figure 4-8. How do we define scale size in this example? Returns to scale is defined for a constant input mix[3] and our measure of scale size must reflect this. For example, in Figure 4-8,

[3] Input mix is used in this thesis to refer to the proportions inputs are to each other.

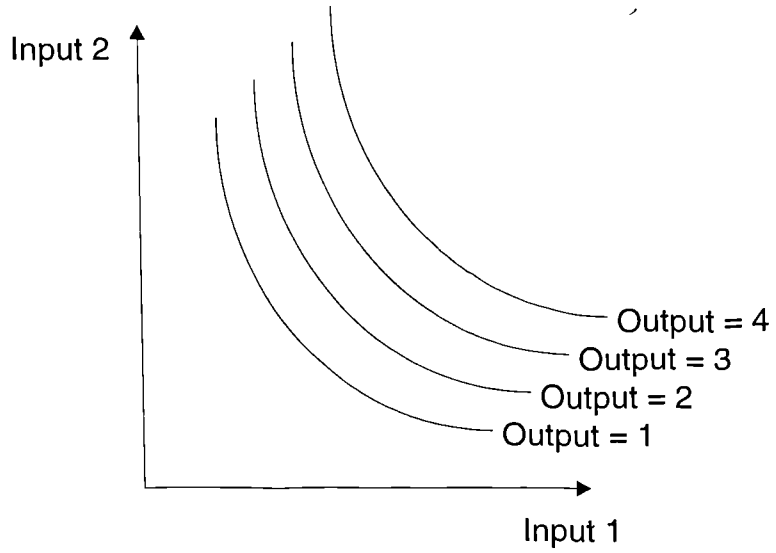
Figure 4-8. Scale size for multiple inputs



DMU B is operating at the same input mix as other DMUs on the radial line from the origin through B, OB. The scale size is defined along this ray through the origin so that DMUs close to the origin are 'smaller' than DMUs further from the origin. In order to be able to define a measure of scale size across input mixes (call this the **cross-mix scale size** (**cmss**)), the inputs can be aggregated using the **isoquant** as is now elaborated.

A production technology in three dimensions can be represented either by a three dimensional graph showing all the relationships between the inputs and outputs or by plotting the input (output) isoquants in input (output) space. See Figure 4-9.

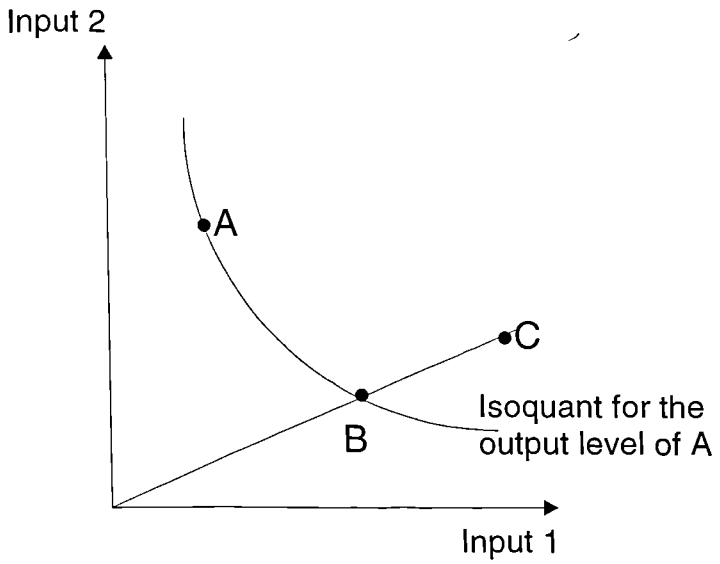
Figure 4-9. Isoquants in input space



The isoquants represent all the efficient input combinations which are capable of producing a certain level of output. For a CRS technology, these isoquants are radial projections of each other and for an increase in all inputs by a factor α, there will be an increase in the output by the same factor. In Figure 4-10, A and C are two units which are on the CRS frontier. Is it possible to say whether A is operating at a scale size that is smaller than, larger than or the same as that of C?

In order to compare the scale sizes of units A and C in Figure 4-10, first compare the scale sizes of unit C and unit B. Unit B is an efficient unit with the same output level as A and the same input mix as C. We have shown that scale size can easily be defined along a specific input mix,

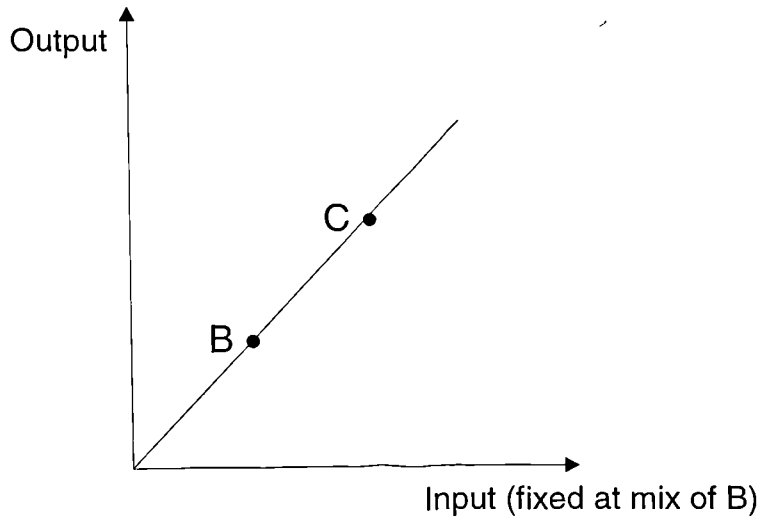
Figure 4-10. Cross-mix scale size for efficient DMUs



as, by considering a specific mix of inputs we have the same scenario as the single-input case. B and C in Figure 4-10 are both on the efficient frontier. C is using more of both inputs than B and is producing more output. If a graph is plotted of the output against the input fixed at the mix of B (see Figure 4-11) then we obtain exactly the same problem as the single-input, single-output case in Figure 4-6. So we can say that C is operating at a larger scale size than B.

Once again, the distance function can be used as a relative scale size measure. Let the inputs of B be the vector $\mathbf{x}_B$. Then the inputs of C must be some multiple, say δ, of this vector, $\delta\mathbf{x}_B$.

Figure 4-11. Scale size for a fixed mix



From Figure 4-11,

$$\frac{S(C)}{S(B)} = \frac{x_C}{x_B} = \frac{\delta x_B}{x_B} = \delta = D_I^{CRS}(y_B, x_C) \qquad (4\text{-}10)$$

From equation (4-6) we have

$$\frac{S(A)}{S(B)} = \frac{y_A}{y_B} \qquad (4\text{-}11)$$

for any two efficient units, A and B in the single-output case.

Therefore, for the cmss we shall define units that yield the same efficient output level to have the same scale size irrespective of the input mix.

That is, all units that lie on the same isoquant in input space are defined to have the same cmss. In Figure 4-10, A and B both have the same output level, $y_A = y_B$, which leads to

$$\frac{S(A)}{S(B)} = \frac{y_A}{y_B} = \frac{y_A}{y_A} = 1 \qquad (4\text{-}12)$$

Therefore, the isoquant is being used to aggregate the inputs.

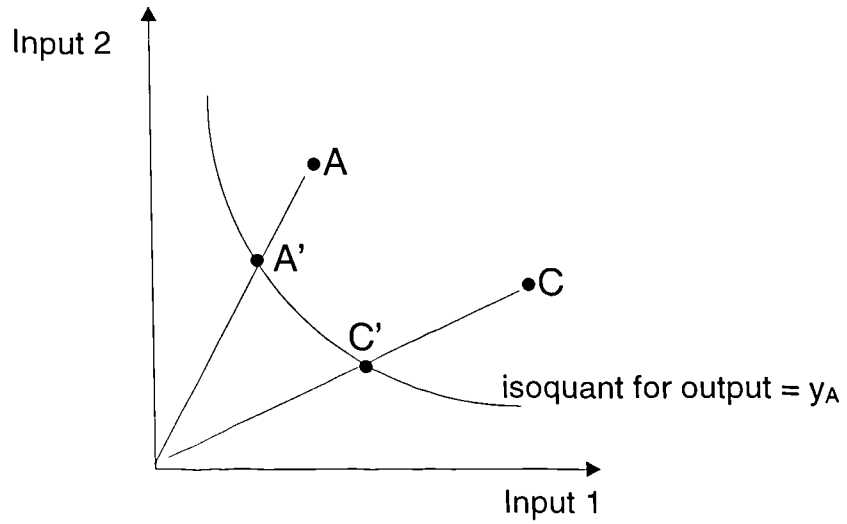
Proposition 3: If DMU A has the same CRS efficient input or output levels as DMU B then the two units are defined to be the same size.

Now the scale size of unit C relative to that of unit A is given by

$$\begin{aligned} \frac{S(C)}{S(A)} &= \frac{S(C)}{S(B)}\frac{S(B)}{S(A)} \\ &= \delta.1 \\ &= \frac{D_I^{CRS}(y_B, x_C)}{D_I^{CRS}(y_A, x_B)} \\ &= \frac{D_I^{CRS}(y_B, x_C)}{D_I^{CRS}(y_B, x_B)} = \delta \end{aligned} \qquad (4\text{-}13)$$

Now neither A nor C need be efficient. We will show that any one of the isoquants can be chosen as a reference for the distance functions used. In the single-input, single-output case, the form of the CRS frontier used to measure scale size did not affect the results - any CRS frontier would give the same relative scale sizes. However, in the multiple input case the shape of the isoquant is important. This shape will be affected by the method used to estimate the production frontier. Assuming that the shape of the isoquant is well estimated by the method then any output level can be chosen to represent the CRS frontier.

Figure 4-12. Cross-mix scale size for inefficient DMUs



Consider two inefficient units, A and C in Figure 4-12. The efficient isoquant for the output level of unit A is shown. The size of unit C

relative to the point on this isoquant with the same input mix as C, C', is given by (c.f. Figure 4-7)

$$\frac{S(C)}{S(C')} = D_I^{CRS}(y_A, x_C) \qquad (4\text{-}14)$$

Similarly, the size of unit A relative to the point on the isoquant with the same input mix, A', is given by

$$\frac{S(A)}{S(A')} = D_I^{CRS}(y_A, x_A) \qquad (4\text{-}15)$$

and the size of A' relative to the size of C' is equal to 1 as point A' and C' lie on the same isoquant. The size of unit C compared to the size of unit A is then given by

$$\frac{S(C)}{S(A)} = \frac{S(C)}{S(C')}\frac{S(C')}{S(A)} = \frac{S(C)}{S(C')}\frac{S(C')}{S(A')}\frac{S(A')}{S(A)} = \frac{D_I^{CRS}(y_A, x_C)}{D_I^{CRS}(y_A, x_A)} \qquad (4\text{-}16)$$

and this is the Malmquist input quantity index.

Now we can measure the relative cross-mix scale sizes of any data set with a single input or output. In order to do this, one unit must arbitrarily be chosen as the reference unit and all other scale sizes can be defined relative to the scale size of this unit.

4.3.3 Calculating the cmss in the single-output case

This chapter has outlined a definition of scale size using the Malmquist quantity index because this will become important when generalising to multiple inputs and outputs. The Malmquist index can be calculated using specially written software or a mathematical programming package.

In practice, the relative cross-mix scale size for the case of a single endogenous output can be measured using the efficient CRS output level rather than the observed inputs (see equation (4-6)). This obviates the need for aggregation of the inputs. The efficient output can be easily calculated by dividing the observed value of the output by the CRS output efficiency. This means that the relative scale sizes for the single output case can be obtained from a normal DEA model and this is a much easier method to implement. Note however, that the Malmquist index will become necessary when we consider multiple-output production in Chapter 6.

In conclusion, in the case of a single endogenous output, the relative cross-mix scale sizes can be calculated using either

(a) **the observed inputs aggregated by the CRS Malmquist input quantity index, or**

(b) **the CRS efficient output level.**

4.4 An example to illustrate how the cmss can be used to identify functional deviation across scale size

This section illustrates the method for defining cross-mix scale sizes for the single-output, two-input case. We will use the data set generated from DGP A as described in Appendix 2. This data set has two inputs and a single output and is piecewise log-linear across the scale size (i.e. piecewise Cobb-Douglas[4]). At small-scale sizes, the production function has IRS moving to CRS at larger scale sizes and finally DRS at the largest scale sizes.

The efficiencies were estimated under DEA using a VRS model and under SF using the Cobb-Douglas function as the estimating function. The Cobb-Douglas function has a fixed value for the scale elasticity so it cannot accommodate increasing, constant and decreasing returns to scale. Therefore, the SF method will exhibit variation of fit across the scale size.

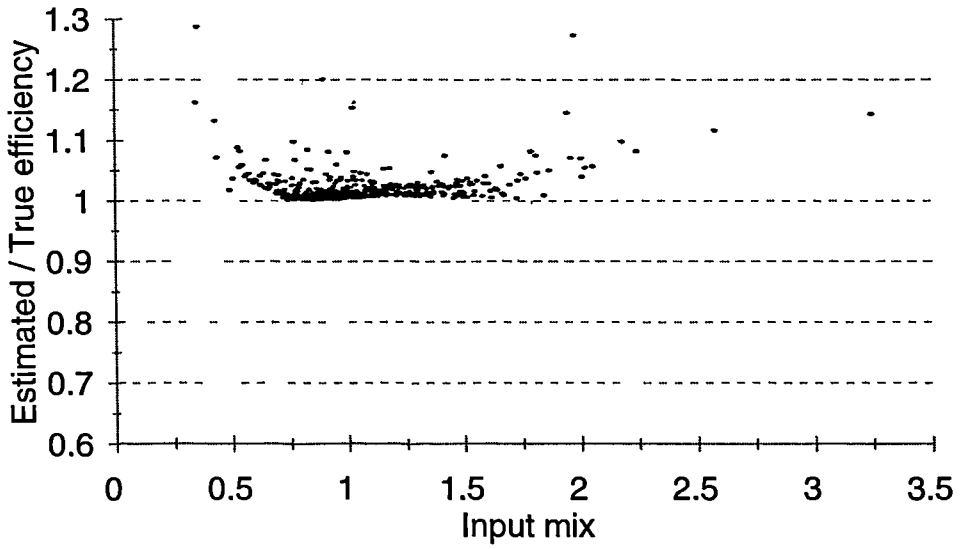
The estimated scale size of each unit was calculated using the CRS efficient output values (observed output/CRS efficiency). Thus the scale size of two units having outputs y_A and y_B and efficiencies e_A and e_B is

[4] Note that this function is non-concave at small-scale sizes so the DEA results at small-scale sizes may be adversely affected as outlined in Hypothesis 7.

$$\frac{S(A)}{S(B)} = \frac{y_A/e_A}{y_B/e_B} \qquad (4\text{-}17)$$

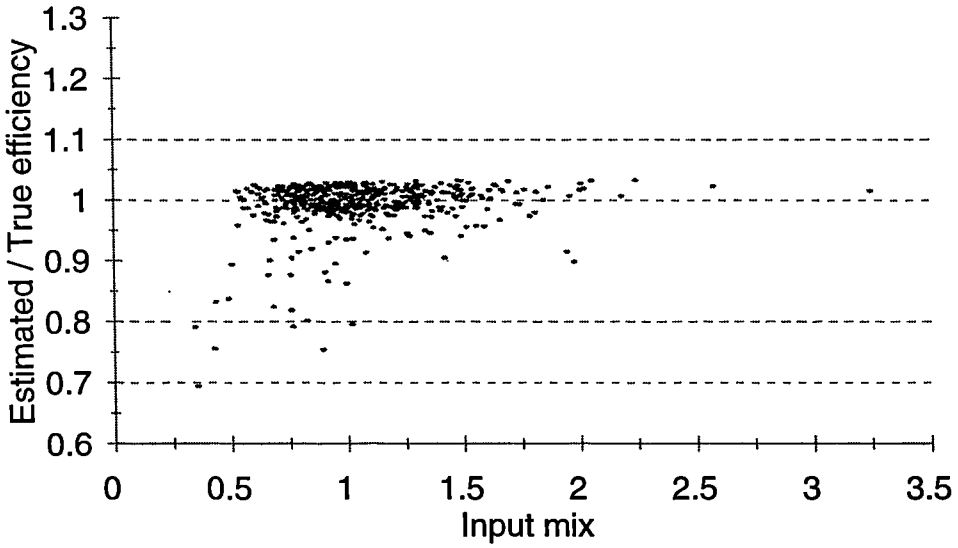
In Figure 4-13 and 4-14, the estimated values from the two methods are plotted across the input mix.

Figure 4-13. A comparison of the performances across input mix: DEA vs True efficiency



In both graphs, the DEA and SF translog methods are showing deviations across the input mix in quite a random manner.

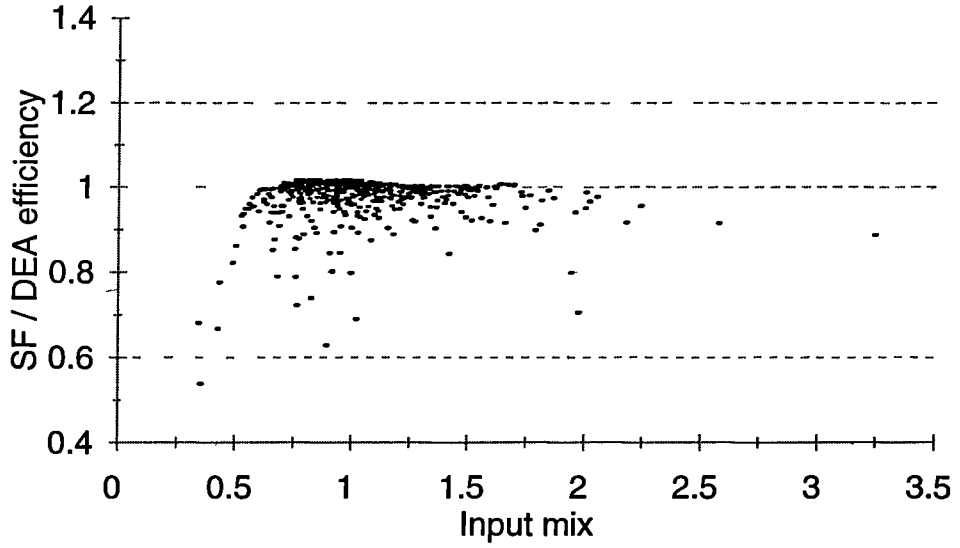
Figure 4-14. A comparison of the performances across input mix: SF vs True efficiency



Similarly, Figure 4-15 compares the ratio of the DEA to SF estimates. It can be seen that there is no obvious pattern to these ratios (except at the very extreme mixes where the SF estimates are greater than those of DEA due to a lack of efficient comparators for DEA at the extreme mixes).

If the difference between the estimated frontier and the true frontier varies as scale size changes or as input mix changes, then a clear relationship should be seen if the functional deviation is plotted against the relevant variable.

Figure 4-15. A comparison of the performances across input mix: DEA vs SF



Figures 4-16, 4-17 and 4-18 give the same ratios plotted across the estimated scale size rather than input mix. In this case there is a much clearer relationship between the estimates and the scale size, particularly for the case of the Cobb-Douglas SF method.

The Cobb-Douglas SF efficiencies deviate from the true efficiencies *only* across scale size because the elasticity of substitution between the inputs is fixed at one, which is, of course, the same as the underlying technology in this case. Therefore, there is no deviation across input mix. DEA is using the data to estimate the elasticity of substitution and this is why there is some deviation across the input mix as well as the scale size.

Figure 4-16. A comparison of the performances across scale size (estimated under DEA): DEA vs True

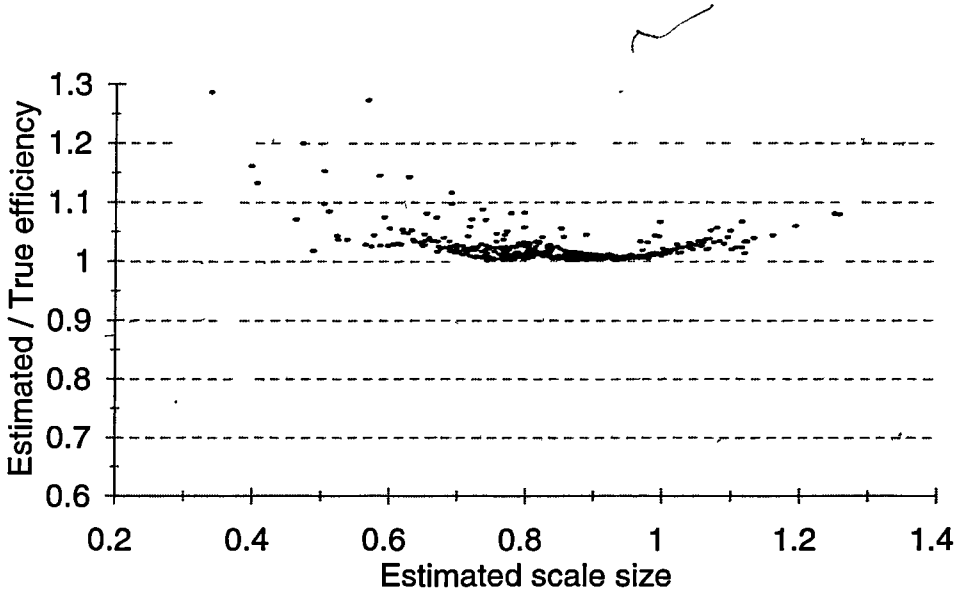


Figure 4-17. A comparison of the performances across scale size (estimated under DEA): SF vs True

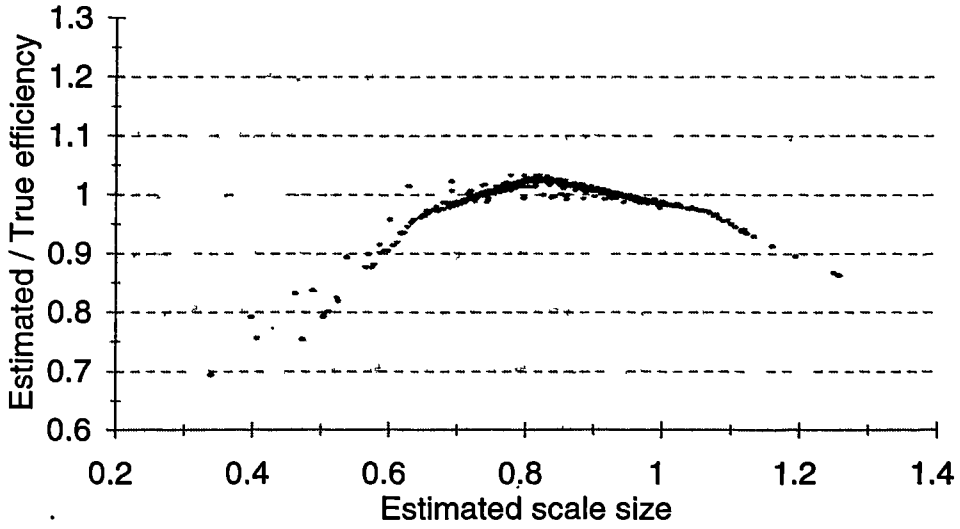
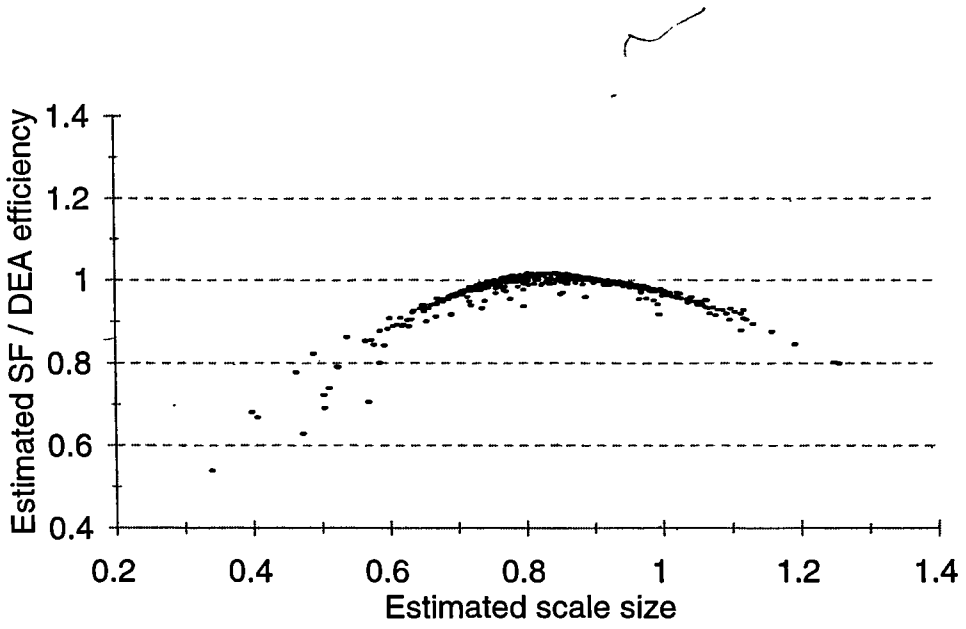


Figure 4-18. A comparison of the performances across scale size (estimated under DEA): DEA vs SF



4.4.1 Most productive scale size

Note that the mpss is still mix dependent. Once each unit has been allocated a cross-mix scale size, by choosing a reference unit (i.e. output level), there is not a specific value or range of values which necessarily gives the most productive scale size. For each input mix there will be a definite range of scale size which is the mpss. However, this range of values will be independent of the mix only for an homothetic technology.[5]

---

[5]An homothetic technology with a single output, satisfies

$$f(\mathbf{x}) = f(\mathbf{x}') \Leftrightarrow f(\mu\mathbf{x}) = f(\mu\mathbf{x}')$$

for any $\mu \in \Re_+$. This means that the 'shape' of the isoquant is constant across the technology. A CRS function with a single output is a special case of an homothetic technology.

4.5 Conclusions

In this chapter the cross-mix scale size has been defined for the single-output, multiple-input case. We have shown that the definition of scale size for units which are not on the frontier depends on the chosen orientation, i.e. whether the input or output is exogenous. This explains why, in an output orientation, the single observed output cannot be used as a measure of scale size. In an output orientation, the input CRS Malmquist index, or the CRS efficient output can be used to measure the scale size of the units and vice versa for an input orientation.

Now that a measure of scale size has been developed in the single output case, this can be used to investigate how the methods perform in different regions of scale size. In the next chapter, the technology will be split into different regions of scale size to illustrate some of the hypotheses which were given in Chapter 2. The three hypotheses which will be tested relate to restrictions on the nature of returns to scale in each method. In the case of DEA, tests will be developed to improve the identification of returns to scale in DEA.

Chapter 5

Using variation of fit to better identify the true nature of returns to scale in DEA

5.1 Introduction

DEA has been used in a very wide range of applications to measure the efficiency of DMUs. In these applications the choice of the CRS[1] or the VRS[2] model often seems to be rather subjective. There will be regions where the frontiers that are formed are very close in the CRS and VRS models. However, for other regions of scale size, the frontiers may vary substantially across the two models yielding very different results for units operating in these regions. In such cases, these DMUs will be given as more - maybe much more - efficient under the VRS model than the CRS model and can be greatly discriminated for or against if the wrong choice of model is made.

In Chapter 2, three hypotheses were put forward for the effect on the results of the methods if the true nature of the returns to scale is not identified. In the first case, Hypothesis 3, a too restrictive assumption about the returns to scale is imposed in the DEA model.

Hypothesis 3

If a restrictive assumption of returns to scale is imposed on the methods unnecessarily, then the estimates will be such that $E_{TRUE} > E_{DEA}$ in the regions where the assumption does not hold. For an homothetic frontier, these regions will vary *only* across scale size.

---

[1] The CRS model (Charnes, Cooper and Rhodes (1978) is Model 1 in Chapter 1.
[2] The VRS model can be found in Banker, Charnes and Cooper (1984).

For example, assume that a CRS frontier is imposed on a VRS, NIRS or NDRS technology. Here we would expect the DEA efficiencies in the non-CRS regions to be less than their true values.

In the case of the SF method, imposing a CRS function on a non-CRS technology is an example of functional misspecification. So, the estimated function will have regions where it lies above and below the true function as argued in Section 2.4.6 in Chapter 2.

Hypothesis 6

If the true frontier is not well estimated by the SF function, then the estimated efficiencies will have regions where they are greater than and less than the true efficiencies across scale size or input mix depending on whether the misspecification varies across scale size or input mix.

In the second case, Hypothesis 4, no restriction is imposed on the returns to scale of the frontier. In the case of DEA this can allow too much 'fit' to the data - there is a danger of units being given efficiency estimates that are too high.

Hypothesis 4

If the underlying technology has CRS, NIRS or NDRS and a less restrictive assumption is imposed on the estimating methods, then the

estimates will be such that $E_{DEA} > E_{TRUE}$ in the regions where the assumption does not hold. These regions will vary across scale size.

In this chapter we will investigate these three hypotheses using 10 simulated sets of data generated according to Data Generating Process B given in Appendix 2. Each data set has 2 inputs and 1 output and is generated from the same **NIRS technology**.

As well as demonstrating these hypotheses, this chapter will propose tests which can be used to identify more accurately the true nature of the returns to scale of the frontier in DEA.

Banker (1996) proposes tests for CRS, NDRS and NIRS across the full range of scale sizes covered by the DMUs being assessed. These tests involve obtaining the inefficiencies under the DEA CRS, VRS, NDRS and NIRS models for all DMUs. For each pair of models, a single test encompassing all scale sizes is then carried out for the closeness of the frontiers. If the two frontiers being tested, e.g. the CRS and VRS frontiers, are close for most of the DMUs then the test will not identify a smaller region where the frontiers do differ. One way to identify such regions is to conduct separate tests for the closeness of the VRS and CRS frontiers for different ranges of scale size. Such tests are developed in this chapter.

A key stage in the development of these tests is the identification of ranges of scale sizes. Chapter 4 developed a measure of scale size for the case where DMUs have a single output and scale size is measured on multiple inputs or vice versa. The measure of scale size developed makes it possible to identify regions of scale size so that the proximity of the CRS and the VRS frontiers within different regions can be tested.

The chapter is structured as follows: The next section explains how possible misspecification of small regions of the frontier may be identified using the scale efficiency measure. The third section identifies regions of possible variation of fit. The fourth section outlines the hypothesis tests that will be used, and the fifth section gives the results of a Monte-Carlo experiment using these tests and illustrates the hypotheses above. The final section is a summary of the method and conclusions from the experiment.

5.2 Scale efficiency and variation of fit in DEA

5.2.1 Variation of fit across scale size

It is possible that in one region of the technology a small subset of all the DMUs operate under returns to scale which do not conform to the bulk of units. If this difference in the returns to scale is not identified by the estimating model, the frontier in this region will be misspecified. This misspecification can be compared to the concept of 'Variation of Fit' outlined in Chapter 2. This was defined as the variation in the

proximity of the estimating and true functions across the technology. This chapter considers possibilities for variation of fit across scale size under DEA and SF. For DEA it may be the case that a non-CRS frontier is identified for certain scale sizes where a CRS frontier has better fit to the true frontier. This could arise in the VRS case because a VRS frontier envelops the data as closely as possible. If there are not enough truly efficient DMUs in a range of scale sizes, then the data may be enveloped very closely by the VRS model leading to overestimates of efficiencies. Whenever a VRS technology is estimated under DEA it is not possible to tell whether the regions which appear to be VRS are truly VRS or 'apparent' VRS because of a lack of efficient CRS units in that region of the frontier.

5.2.2 A measure of variation of fit

A measure of this variation of fit for the single output case, the functional deviation, was defined in Chapter 2 for DMU j, as:

$$FD_j = \frac{\text{estimated efficient output of DMUj}}{\text{true efficient output of DMUj}}. \qquad (5\text{-}1)$$

This definition is useful in a simulated situation when we know what the true efficient output is. In real applications we will not be able to measure the functional deviation but the scale efficiency will be closely related.

If the CRS frontier is the true frontier and a VRS frontier is estimated, then the functional deviation will be very close to the scale efficiency. For example, in Figure 5-1 the functional deviation of DMU A will be given by

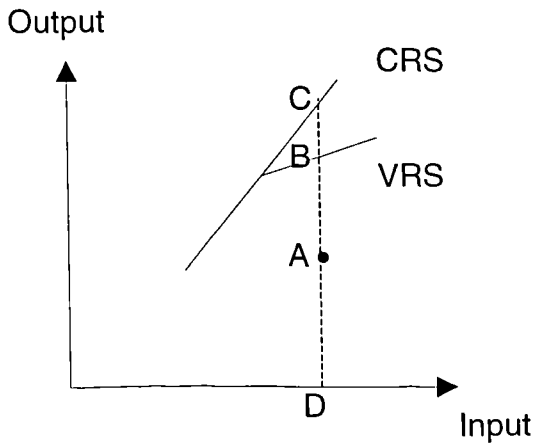
$$FD_A = \frac{\text{estimated efficient output of DMU A}}{\text{true efficient output of DMU A}} = \frac{DB}{DC}. \qquad (5\text{-}2)$$

The scale efficiency of DMU A will also be given by

$$\text{scale efficiency} = \frac{DB}{DC}. \qquad (5\text{-}3)$$

Similarly, if the VRS frontier is the true frontier and a CRS frontier is

Figure 5-1. Functional deviation and scale efficiency

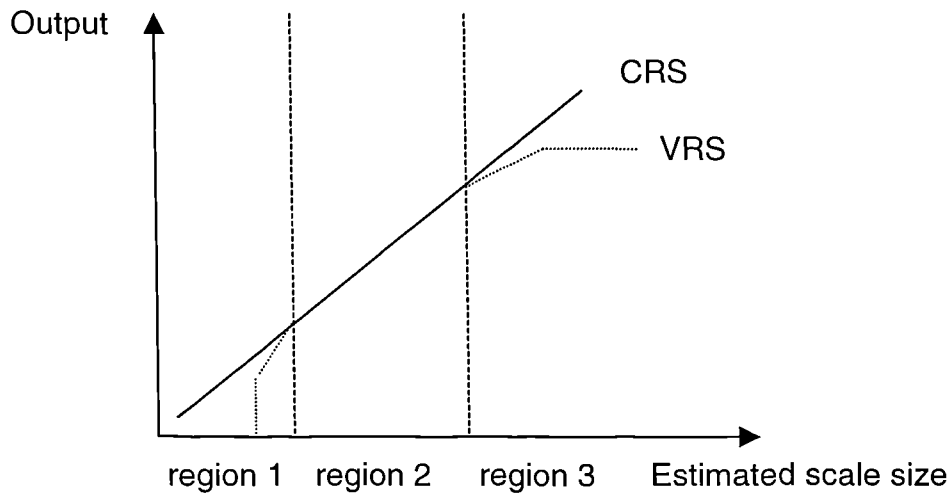


imposed, the functional deviation will be very close to the inverse of the scale efficiency. We will use this relationship between the functional deviation and the scale efficiency as an indicator of possible variation of fit. (Possible, in the sense that if there is no variation of fit, the scale efficiency will be an indicator of true scale efficiency and not functional deviation.) If there are only a few units which have this functional deviation, the Banker (1996) tests would not identify them, and this is why tests for different returns to scale in these regions will be developed here.

5.3 Identifying the regions where variation of fit may occur

As noted earlier, current tests (e.g. Banker (1996)) for differences between the CRS and VRS efficiencies may not detect such differences if they relate to only a small region of scale sizes rather than the entire set of DMUs. Figure 5-2 illustrates the problem which can be encountered for a single-output, multiple-input homothetic production technology where the scale size of each DMU has been identified using the method outlined in the previous chapter.

Figure 5-2. The different regions observed in a DEA analysis



Three regions have been identified. In the scale size range labelled region 2, most DMUs are given the same efficiencies under the CRS and VRS DEA models, i.e. this is an approximation of the most productive scale size. If the bulk of the observed DMUs operate in this region and if their CRS and VRS efficiencies are close, tests such as those by Banker (1996) applied to all DMUs would fail to reject the null hypothesis that the CRS and VRS efficiencies follow the same distribution. Yet there could be substantial ranges of scale size corresponding to region 1 and region 3 where the CRS and VRS frontiers differ greatly. These regions can be identified by plotting the scale efficiency against the scale size of the DMUs as will be illustrated later. This raises an important question. How can we tell, when estimating the efficiencies using DEA, whether the DMUs in regions 1 and 3 are inefficient under CRS or in fact operating under VRS?

The definition of a measure of scale size, when scale is measured on multiple variables, makes it possible to test DMUs such as those operating in regions 1 and 3, against DMUs such as those in region 2, to identify whether they operate under the same nature of returns to scale.

Note that, the frontier from a NIRS DEA model will give the CRS frontier in regions 1 and 2 and the DRS frontier in region 3. Similarly, the NDRS DEA model will give the IRS frontier in region 1 and the CRS frontier in regions 2 and 3.

5.4 Hypothesis testing for returns to scale in DEA

We will use our measure of scale size to identify DMUs which operate in particular ranges of scale size and will then test whether the same type of returns to scale hold as for DMUs operating in other regions of scale size.

We will use five different tests and compare their performance. Three of the tests are those proposed by Banker (1993), the fourth is the Mann-Whitney test and the fifth a test for the difference between means.

The five tests for differences between two groups, G_1 and G_2, of inefficiencies which we will be using are summarised in Appendix 3.

In this chapter we propose to test for the nature of returns to scale holding for DMUs in **different ranges** of scale size rather than for all DMUs. This offers an additional advantage in that the inefficiencies of DMUs in different ranges of scale size are more likely to be independent than would be the case if all DMUs are used.

5.5 A Monte-Carlo simulation

The 10 data sets are all generated as described in Appendix 2, Data Generating Process B. The technology is piecewise log-linear and has CRS for most of the units and DRS for the largest units. Note that this is an NIRS frontier. Each data set has a low random noise and a half-normal inefficiency term.

The DEA CRS, VRS, NIRS, NDRS models were used to compute, for each set of 250 DMUs, the corresponding DEA efficiencies. This gave two DEA estimates of the inefficiency for each DMU (as one of the NIRS and NDRS estimates is the same as the CRS estimate and the other is the same as the VRS estimate). A CRS translog SF model was also estimated for each data set assuming truncated normal distribution for the inefficiency.

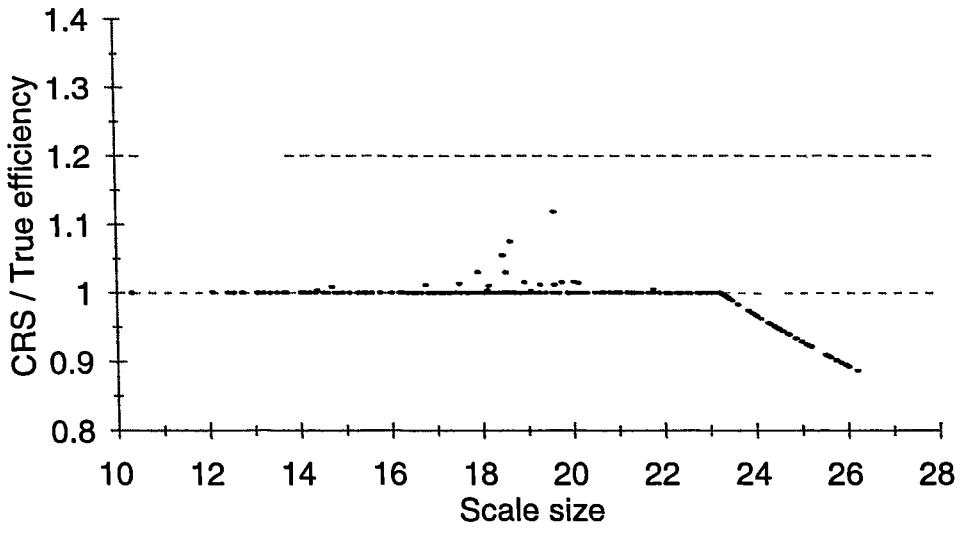
Each of the three hypotheses, 3, 4 and 6, will now be illustrated by considering one of the ten data sets. (In this case, the observed

outputs are generated with no random noise in order to test just one assumption at a time.)

5.5.1 Illustrating Hypothesis 3

Hypothesis 3 states that if a too restrictive assumption of returns to scale is imposed on the data set, the DEA efficiencies will be less than the true efficiencies in the regions where this assumption does not hold. If CRS is imposed on the data sets here (which are truly NIRS) then the region where the frontiers differ is for large scale sizes, which are truly DRS. In this region the efficiencies estimated under the CRS DEA model are likely to be less than the true values.

Figure 5-3. Imposing a CRS frontier on an NIRS data set

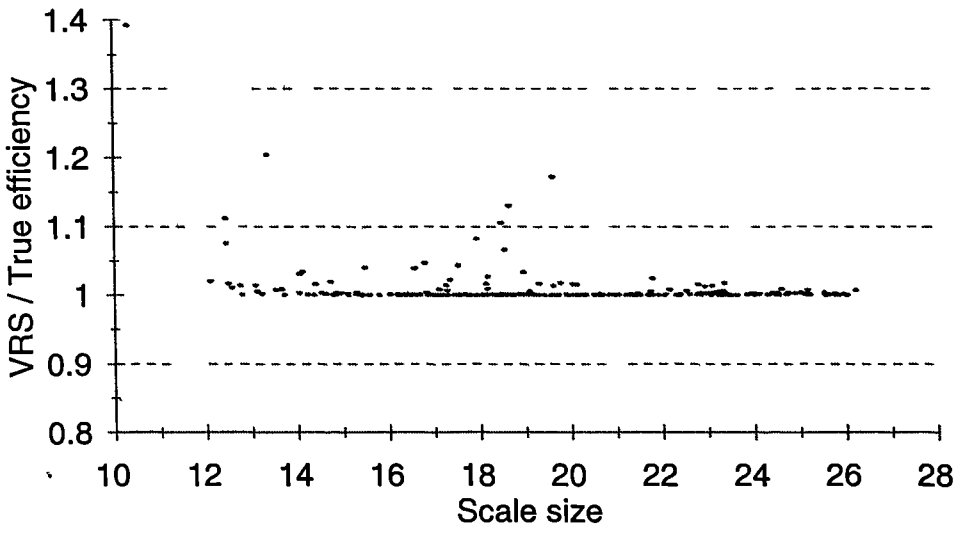


In Figure 5-3, the ratio of the estimated efficiencies to the true efficiencies (the functional deviation) is plotted across scale size (here, the true efficient output). It is clear that there is variation of fit across scale size and that the high scale size units are in a region of poor fit.

5.5.2 Illustrating Hypothesis 4

This data set can also be used to illustrate Hypothesis 4. If the DEA method is not restrictive enough in its assumption of returns to scale, i.e. in this case a VRS model is used, there may be regions where the efficiency estimates are larger than the true values because the form of the frontier is over accommodating. To illustrate this hypothesis, see Figure 5-4 where the functional deviation is again plotted against the true scale size. However, in this case the efficiency estimates are taken from the VRS DEA model.

Figure 5-4. Allowing for a full VRS DEA frontier

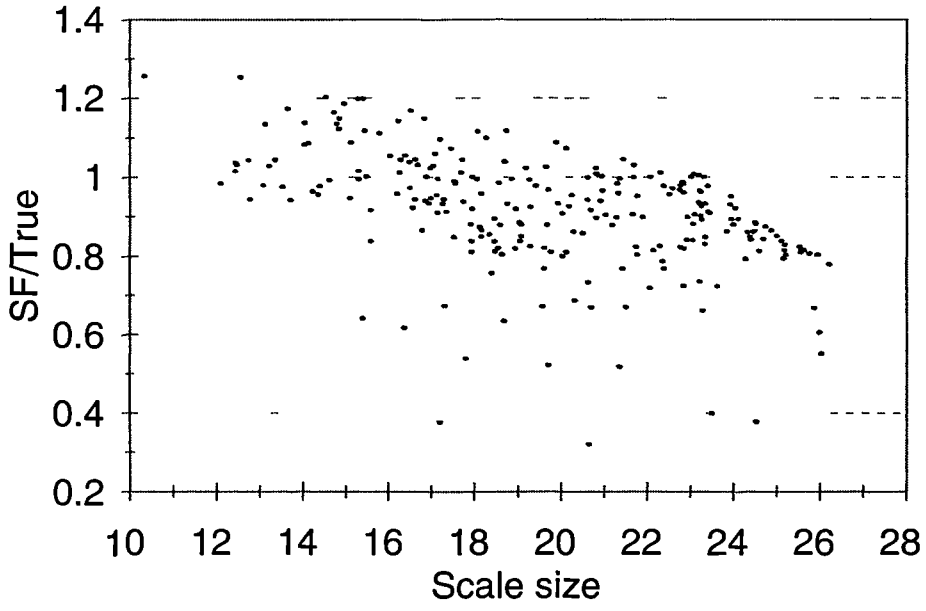


In this case, it is again clear that there is variation of fit across scale size. Here it is the units at the medium and small-scale sizes which are in a region of poor fit.

5.5.3 Illustrating Hypothesis 6

If the SF method imposes CRS on the technology, and in this case the true nature of returns to scale is NIRS, then according to Hypothesis 6, we expect regions where the SF estimates are greater than the true efficiency values and regions where they are less. The proximity of the estimated function to the true function should vary across scale size.

Figure 5-5. The SF estimates compared to the true values across scale size



In Figure 5-5, the ratio of the SF efficiency estimate to the true efficiency estimate for each unit is plotted against the scale size.

It is clear from this graph that there are definite regions where the estimated frontier lies above the true frontier (scale sizes above 23) and other regions where it lies below (scale sizes below about 15). This leads to the efficiency estimates being greater than the true values in some regions and less than the true values in other regions.

5.5.4 Testing for returns to scale across the full range of scale sizes

Tests 1 to 5 outlined in the previous section were used to test whether the full range of DMUs in each data set generated operates under CRS, VRS, NDRS or NIRS. The results are given in Table 5-1.

The table details the number of times out of 10 that the null hypothesis, indicated at the column heading, is rejected at the 5%, or in brackets, at the 10% significance level. For example, the entry 3 (6) in the column headed CRS means that out of the 10 tests for CRS vs. VRS, 3 are rejected at the 5% level and 6 are rejected at the 10% level under Test 1.

Note that the small-scale units are truly operating under CRS and the large-scale units are truly operating under DRS. The results in the CRS

Table 5-1. Testing for CRS, NIRS and NDRS across the whole technology

	Null hypothesis:	CRS	NDRS[a]	NDRS[b]	NIRS[a]	NIRS[b]
	Alternative hypothesis:	VRS	DRS[$]	DRS[$]	IRS[$$]	IRS[$$]
Test 1	Assuming exponential	3 (6)	0 (0)	0 (0)	0 (0)	0 (0)
Test 2	Assuming half-normal	1 (4)	0 (0)	0 (0)	0 (0)	0 (0)
Test 3	Kolmogorov-Smirnov	8 (9)	0 (0)	0 (0)	1 (1)	1 (1)
Test 4	Mann-Whitney	9 (10)	0 (4)	0 (6)	1 (5)	1 (4)
Test 5	Difference in Means	8 (9)	0 (1)	0 (1)	0 (2)	1 (2)

The numbers in the table indicate the number of times out of 10 that the null hypothesis was rejected at the 5% (10%) level.

NDRS[a] compares the VRS and NDRS inefficiencies[3].

NDRS[b] compares the CRS and NIRS inefficiencies[3].

[$] Rejection of the NDRS[1] or NDRS[2] null hypothesis leads to the conclusion that there is a region of the technology which operates under DRS.

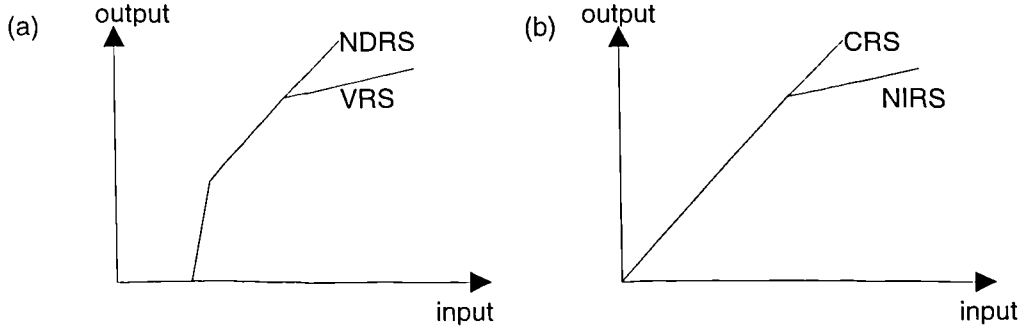
NIRS[a] compares the VRS and NIRS inefficiencies.

NIRS[b] compares the CRS and NDRS inefficiencies.

[$$] Rejection of the NIRS[1] or NIRS[2] null hypothesis leads to the conclusion that there is a region of the technology which operates under IRS.

[3] Note that there are two ways to test the null hypothesis of NDRS: the first is to compare the VRS and NDRS efficiencies, see Figure 5-6(a), and the second is to compare the CRS and NIRS efficiencies, see Figure 5-6(b). Similarly for NIRS.

Figure 5-6. Testing the null hypothesis of a NDRS frontier



column of Table 5-1 would appear to suggest that the DMUs do not operate under CRS but beyond this it is difficult to make further conclusions. Neither the NIRS nor the NDRS null hypotheses are rejected in general. Under these circumstances, if a VRS DEA model is decided upon, all the DMUs reflected onto the IRS frontier will be given over-estimates of their true efficiencies.

The next section shows how, by testing DMUs operating within different ranges of scale size, it is possible to identify more accurately the nature of the returns to scale.

5.5.5 Calculating the relative cross-mix scale sizes

Ranges of scale size were identified after first determining the relative cross-mix scale size of each DMU in each one of the 10 data sets generated.

The relative cross-mix scale size (cmss) was calculated using the CRS DEA efficient output. (Note that the Malmquist input distance function could have been used instead and would have given exactly the same relative scale sizes.) The CRS efficient output was calculated by multiplying the observed output by the CRS DEA output efficiency. The relative cmss of each DMU was then given by taking one of the DMUs as the reference unit and calculating the ratios

$$\frac{S(DMU_j)}{S(DMU_r)} = \frac{y_j}{y_r}, j = 1,...,n \qquad (5\text{-}4)$$

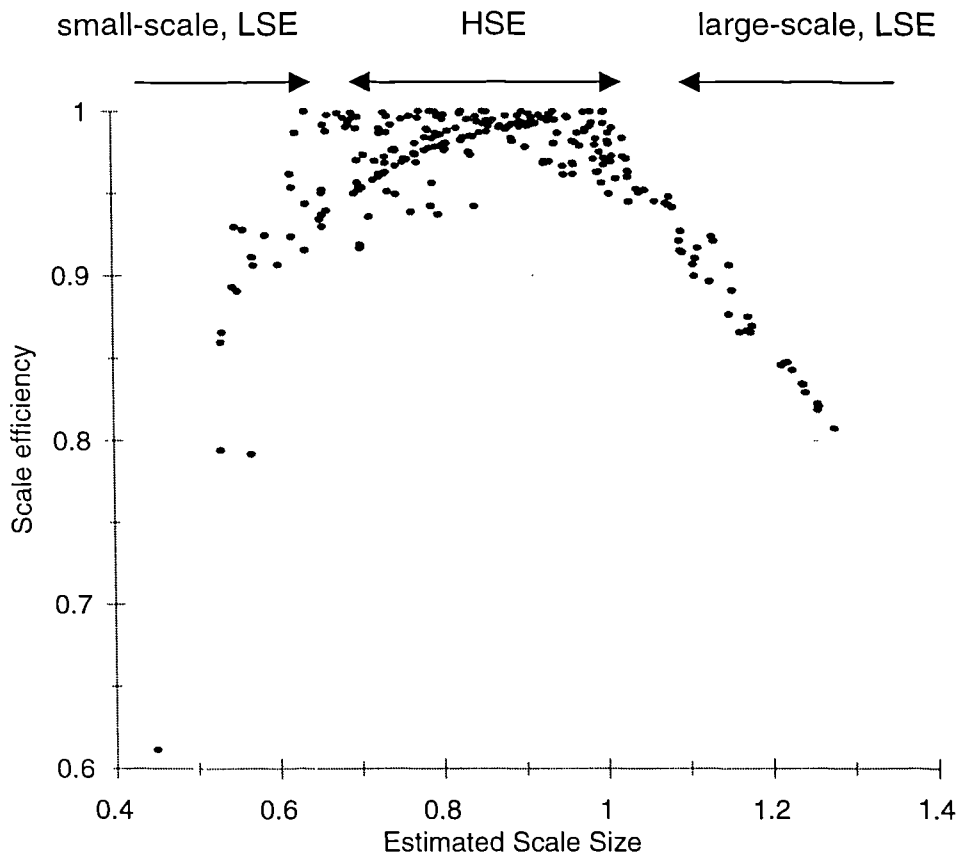
where DMU_r is the reference DMU and y_j is the CRS efficient output for DMU_j. (Note that if DMU_r is a unit with output = 1, the relative scale sizes would be given by the CRS efficient output.)

5.5.6 Testing for returns to scale across specific ranges of scale sizes

Once the relative scale sizes have been calculated, the scale efficiency can be plotted against the estimated scale size as an indicator of possible variation of fit (as outlined in the first section). This is done in Figure 5-7 for one of the data sets for illustrative purposes.

It will be possible to plot a similar graph no matter how many inputs and this graph can be used to identify whether there is any possibility of variation of fit.

Figure 5-7. A graph showing scale efficiency across scale size as an indicator of possible variation of fit



Firstly, the graph should be inspected to see if a pattern can be subjectively identified - i.e. ranges of scale size where scale efficiencies are high and regions where they are low. If none can be identified then there is nothing to test. If there is a pattern, the different ranges of scale sizes where returns to scale appear to vary should be identified. In Figure 5-7 we can discern a range of scale sizes where scale efficiencies are generally high (stretching from about scale size 0.7 to 1) and two scale sizes (one small, up to about 0.7, and one large, over 1)

where scale efficiencies are low. We shall refer to these ranges as 'HSE', 'small scale, LSE' and 'large scale, LSE' respectively, where HSE stands for 'high scale efficiency' and LSE for 'low scale efficiency'.

In the general case, the pattern of scale inefficiency variation may not be the same as that in Figure 5-7, nor as clear, but the different ranges of scale size should be distinguishable. If there is no obvious pattern to the graph, it is likely to be because the production function is not homothetic and the nature of returns to scale is varying across the input mix as well as across the scale size (or there is no variation of fit).

Next, the groups of DMUs which we will test need to be created. Rather than taking the whole set of DMUs within each of the ranges identified, we choose to take a subset of them. This should reduce any bias created by the choice of the cut-off point for the range. We create three sets of DMUs as follows:

The first set consists of the smallest scale size DMUs taken from the 'small scale, LSE' range identified earlier. The size of the set is subjective but it is suggested it be 10% of all DMUs, provided that this results in a sample size to make the test(s) to be used statistically valid and the DMUs do come from the range identified. The second set consists of 10% of all DMUs this time drawn from HSE DMUs. We will use this as the reference set for the tests - we assume that these units

have estimated efficiencies which are very close to their true efficiencies. The third set consists of the largest 10% of all DMUs taken from the 'high scale size, LSE' group.

5.5.7 Testing by region

Table 5-2 gives the results of the five tests outlined earlier for testing for the nature of the returns to scale holding at different ranges of scale size. From these results it is possible to see how the tests perform for the 10 sample data sets. In each case the null hypothesis is that DMUs in the HSE range and the LSE range concerned operate under the same type of returns to scale. The returns to scale tested are identified in the column heading.

Table 5-2. Testing the nature of returns to scale by region

		Small scale sizes		Large scale sizes	
	Null hypothesis:	local CRS	local IRS	local CRS	local DRS
Test 1	Assuming exponential	0 (0)	9 (10)	2 (7)	0 (0)
Test 2	Assuming half-normal	0 (0)	0 (10)	5 (7)	0 (0)
Test 3	Kolmogorov-Smirnov	0 (0)	10 (10)	8 (10)	0 (0)
Test 4	Mann-Whitney	0 (0)	10 (10)	10 (10)	0 (0)
Test 5	Difference in means	0 (0)	8 (10)	9 (10)	0 (0)

Table 5-2 gives the number of times out of 10 that the null hypothesis is rejected. In each case, the alternative hypothesis is that the two groups of units do not have the same type of returns to scale.

The groups of DMUs used in the tests have now been constructed so that the inefficiencies of the DMUs in the 'small scale size' and 'large scale size' LSE regions are tested against the inefficiencies in the HSE region.

It is recalled that, by construction, the small-scale size LSE DMUs operate as the HSE DMUs, i.e. under CRS. In contrast, the large scale size LSE DMUs operate under DRS. It can be seen from Table 5-2 that all the tests perform reasonably well - those that do not assume a distribution for the inefficiency term performing better than those that do. Small scale size DMUs being CRS is not rejected, but being VRS is rejected by (almost) all tests and these are correct outcomes given the technology in (A2-4). For large-scale size LSE DMUs, the CRS hypothesis is correctly rejected very frequently, and the VRS is always accepted - consistent with the large-scale size DMUs operating under DRS. On balance, the results from this method of splitting the units into regions by scale size has given significantly better results than the tests across the full set of DMUs detailed in Table 5-1.

It should be noted that the inefficiencies used in the tests by region developed in this chapter are more likely to be independent than the inefficiencies of all DMUs tested by assumption of returns to scale. This is because the inefficiencies of a DMU under alternative assumptions of returns to scale are likely to be correlated.

5.6 Conclusions

This chapter has addressed the issue of returns to scale differing over certain ranges of scale size and our ability to identify such ranges. It has been shown how an incorrect assumption about the nature of the returns to scale in the DEA method can affect the efficiencies of units in specific regions of scale size by illustrating Hypotheses 3 and 4 from Chapter 2. An illustration of Hypothesis 5 also showed how the SF results could be affected by functional misspecification due to too restrictive an assumption about the nature of the returns to scale. It was then shown how the measure of scale size proposed in the previous chapter makes it possible to group DMUs by scale size and test for the nature of returns to scale at different scale sizes.

In order to carry out the tests, the data set needs to be sorted according to scale size. Once this unique measure of scale size has been estimated, it is possible to construct graphs such as Figure 5-7 which give intuitive support as to where possible variation of fit may occur.

These regions can then be tested to see whether the local VRS frontier identified is real or 'apparent' due to a lack of efficient units.

A simulation was used to illustrate how the new tests by scale size ranges better identify the true nature of returns to scale. The tests are useful in practice because the true underlying technology for a given data set is never known, so before an efficient frontier can be estimated a decision must always be made as to which DEA model should be used. As tests across all scale sizes are weighted heavily by the majority of DMUs they can fail to detect small sets of DMUs that operate under returns to scale which differ from those of the main body of DMUs. By taking only the DMUs in the regions where we think variation of fit may occur it is possible to eliminate this weighting and greatly improve the results.

Identifying the true nature of returns to scale has important policy implications. In an efficiency analysis, the information gained can be used to inform future strategy. If a DMU is found to be operating under a region of the frontier which exhibits IRS then it would be worthwhile enlarging the operation of that unit. That is, for each extra 1% of input used, more than an extra 1% of output will be produced. If this region of the frontier has been misclassified then these economies of scale may not hold and there may be no benefits in increasing the scale of operation.

In the next chapter the definition of cross-mix scale size will be generalised to the multiple-output case.

Chapter 6

Measuring cross-mix scale size

in multiple dimensions

6.1 Introduction

Having outlined a method for comparing the efficiency estimates of two methods across scale size for the case of a single input or output, this chapter moves on to consider the most general case of multiple inputs and outputs. In order to be able to investigate how the efficiency estimates of the methods compare across scale size in multiple dimensions, it is necessary firstly to be able to implement both methods in multiple dimensions and secondly, to define a relative cross-mix scale size in multiple dimensions.

DEA easily generalises to multiple dimensions which is often cited as its biggest advantage over parametric methods such as SF which are generally restricted to a single input or output. However, various methods can be used to define multi-input, multi-output functions and these will be discussed in Sections 6.2, 6.3 and 6.4 of this chapter. Section 6.5 generalises the method for defining a relative cross-mix scale size to multiple outputs. Section 6.6 gives an example to show how this method can be used in practice.

6.2 Defining parametric production frontiers in multiple dimensions

So far only frontiers which have a single input or output have been considered[1].

e.g.

$$\ln y = f(\beta;\ln x) + v - u. \qquad (6\text{-}1)$$

In order to analyse data sets which have multiple inputs and outputs, (6-1) must be generalised to include multiple outputs. This is not straightforward. There are, however, several approaches to defining production frontiers in the multi-input, multi-output case.

The production frontier can be defined implicitly as

$$F(x,y) = 0 \qquad (6\text{-}2)$$

where F is assumed to be continuously differentiable, strictly increasing in y and strictly decreasing in x.

In Caves, Christensen and Diewert (1982), a multiple-output production frontier is specified as

---

[1] Refer to chapter 1 for review of SF in single output, multi-input technology.

$$y_1 = f(y^*,x), \qquad (6\text{-}3)$$

where y_1 is one of the outputs and y^* is the vector of outputs excluding y_1. By specifying the frontier in this manner (see also Samuelson (1966)), the error is assumed to affect only one of the outputs and all the other outputs are assumed to be independent of the inputs, and exogenously fixed. This may be reasonable if all but one of the outputs are uncontrollable. However, this is generally not the case.

6.2.1 Price data is available

The standard approach used to define multiple output technologies is to estimate a cost, revenue or profit function instead of the production function. This is known as Duality theory (Shephard (1970), McFadden (1978), Lau (1972), Jorgenson and Lau (1974), Färe and Primont (1995) and (1996)). However, in each of these cases information is required about the input or output prices or both.

If price data is available for

- the inputs, it is possible to specify a cost function (assuming cost minimisation);
- the outputs, a revenue function can be specified (assuming revenue maximisation);

- both the inputs and outputs, a profit function can be specified (assuming profit maximisation).

6.2.2 Price data is not available

DEA has been used widely in the not-for-profit sector where little or no price data is available. How can SF methods be used in this situation? The production relationship between multiple inputs and multiple outputs needs to be specified parametrically.

Two different methods have been developed to handle production functions in multiple dimensions:

- Distance functions

  Distance functions were introduced in Chapter 4 to measure the relative cross-mix scale sizes of a set of units. See the next section for a generalisation to multiple dimensions.

- Stochastic ray production frontier

  The stochastic ray frontier, developed by Löthgren (1997), is a generalisation of the single output SF model. This method writes the Euclidean norm of the output vector as a function of the output mix and the inputs. See Section 6.4 of this chapter for a full explanation.

6.3 Distance functions in multiple dimensions

The distance function was introduced in Chapter 4 to measure the cross-mix scale size in the single-output case. As it is equivalent to the inverse of the technical efficiency measure, the distance function can be used to measure efficiency in multiple dimensions. The distance function is an alternative way to specify the production function. When the distance function is equal to 1, the unit must be lying on the production function. Therefore, instead of specifying the production function as the maximum output that can be achieved from given inputs, the distance function can specify the frontier as all the points which are technically efficient, i.e.

$$D(x,y) = 1. \qquad (6\text{-}4)$$

The stochastic distance function can be defined parametrically by

$$1 = D(x,y)\exp(v - u). \qquad (6\text{-}5)$$

where D(x,y) is a suitably flexible functional form. When there is no inefficiency or random noise the distance function is equal to 1 and as the inefficiency increases, the distance function decreases.

The stochastic distance function is difficult to estimate parametrically as the specification involves no dependent variable on the left-hand side of

the equation. However, the distance function is equivalent to the stochastic ray production frontier as shown in Löthgren (1997), and as the ray production function is easier to estimate, this is the method which will be used here and is explained in the following section.

6.4 The stochastic ray production frontier

The stochastic ray production frontier (Löthgren (1997)) generalises the single-output stochastic frontier to multiple outputs by converting the output vector to polar co-ordinates.

For the single output case, the technology is described by the output set

$$P(x) = \{ y \in \Re_+ : x \text{ can produce } y \} \qquad (6\text{-}6)$$

and the production function is given by

$$f(x) = \max \{ y \in \Re_+ : y \in P(x) \} \qquad (6\text{-}7).$$

In multiple dimensions the technology can similarly be described by the output set

$$P(x) = \{ y \in \Re_+^s : x \text{ can produce } y \} \qquad (6\text{-}8)$$

It is obviously not straightforward to generalise the production function to multiple outputs. In order to define the multiple-output production function, rewrite the output vector in polar co-ordinates:

$$y = r \cdot m(\theta) \qquad (6\text{-}9)$$

where $r = \|\mathbf{y}\|$, the Euclidean norm of the output vector, and

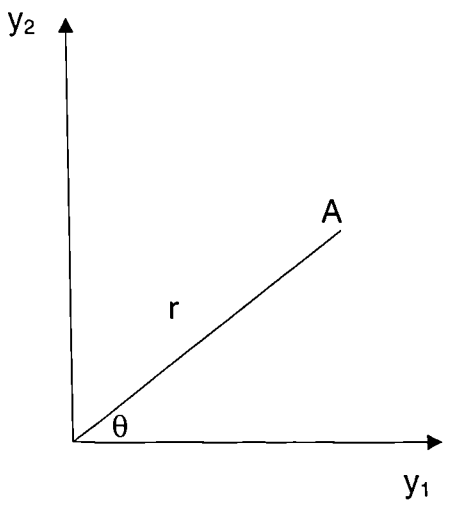
$$m_i(\theta) = \cos(\theta_i)\prod_{j=0}^{i-1}\sin(\theta_j), \text{ for } i = 1,\ldots,s$$

$$\sin(\theta_0) = \cos(\theta_s) = 1,\ \theta \in [0,\pi/2]^{s-1}.$$

The polar co-ordinate angles θ_i are obtained recursively.

For example, consider the 2-output case shown in Figure 6-1.

Figure 6-1. Polar co-ordinates in 2 dimensions



The y_1 co-ordinate is given by $r \cdot \cos(\theta)$ and the y_2 co-ordinate by $r \cdot \sin(\theta)$.

Therefore, $m_1(\theta) = \cos(\theta)$, $m_2(\theta) = \sin(\theta)$.

From the method above, $\mathbf{y} = r \cdot m(\theta)$ and $m_i(\theta) = \cos(\theta_i) \prod_{j=0}^{i-1} \sin(\theta_j)$, for

$i=1,2$, $\sin(\theta_0) = \cos(\theta_2) = 1$, $\theta \in [0,\pi/2]$.

So $m_1(\theta) = \cos(\theta_1)\sin(\theta_0) = \cos(\theta_1)$ and

$m_2(\theta) = \cos(\theta_2)\sin(\theta_0)\sin(\theta_1) = \sin(\theta_1)$.

The multiple output production function can now be written as

$$f(x, \theta) = \max \{ r \in \Re_+ : r \cdot m(\theta) \in P(x) \} \qquad (6\text{-}10)$$

This gives the maximum norm of the attainable outputs given inputs x and output mix represented by θ. This formulation does not suffer from the problems of causality of the formulation given in (6-3). It is assumed that the output *mix* is exogenously fixed rather than all but one of the outputs.

In order to estimate the stochastic frontier, a combined error term is added:

$$r = f(x,\theta) \exp(v-u) \qquad (6\text{-}11)$$

or

$$\ln r = g(\ln x, \ln \theta) + v - u \qquad (6\text{-}12)$$

where e^{v} is the random error term and e^{-u} is the inefficiency term. This leads to a radial measure of efficiency in the same sense as the DEA measure in Model 1. In order for an inefficient unit to reach the frontier, all outputs must be increased by the same factor.

This frontier can be estimated using any software suitable for estimating stochastic frontiers.

6.5 Measuring cross-mix scale size in multiple dimensions

Now that several methods have been outlined for specifying a multiple-input and output production function, a method must be developed for measuring the cross-mix scale size of the DMUs when the technology has multiple inputs and outputs.

In the single output case, only after units have been projected onto the same CRS frontier can the relative sizes of the units be discussed. Therefore, the choice of orientation is important in defining the size of a unit. In the case of multiple inputs and outputs this choice should again

be quite straightforward, as generally, either the inputs or outputs can be considered exogenous.

From Chapter 4 we have three propositions:

Proposition 1: If DMU A has observed output and input levels which are all greater than DMU B then DMU A must have a larger cmss than DMU B.

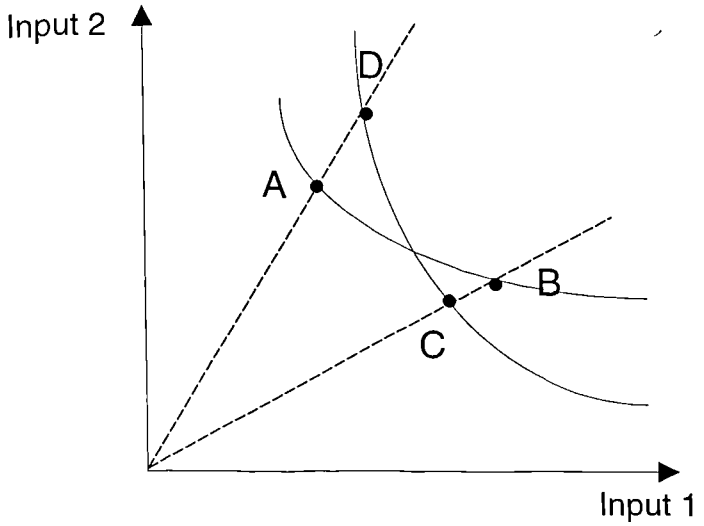
Proposition 2: If DMU A has CRS efficient output (or input) levels which are all greater than those of DMU B then DMU A must have a larger cmss than DMU B.

Proposition 3: If DMU A has the same CRS efficient output or input levels as DMU B then the two units will be said to have the same cmss.

In the multiple-input, multiple-output case this does not cover all eventualities, as it is also possible for unit A to have some inputs which are smaller than those of unit B, and some which are larger *and* some outputs which are smaller than those of unit B and some which are larger. For this case, we cannot say whether the size of A is larger, smaller or the same size as DMU B without aggregating the inputs or outputs.

In multiple dimensions, the inputs or outputs cannot be aggregated by the normal CRS frontier.

Figure 6-2. CRS isoquants in multiple dimensions



In multiple dimensions (i.e. more than one input *and* more than one output), it is possible for CRS isoquants to cross. This is due to the fact that each output vector defines an isoquant of vectors in input space. Consider the technology shown in Figure 6-2. The two isoquants shown in the diagram are both CRS isoquants.

Point B has the same input mix as C and the same output vector as A. B and C cannot lie on the same isoquant because all B's inputs are larger than C's so if B and C have the same output levels, then B would be inefficient. Therefore, from proposition 1, B must have a larger cross-mix scale size than C:

$$S(B) > S(C). \qquad (6\text{-}13)$$

B lies on the same isoquant as A, therefore, B and A can produce the same output levels and from Proposition 3, B and A can be said to have the same cross-mix scale size

$$S(B) = S(A) \Rightarrow S(A) > S(C). \qquad (6\text{-}14)$$

In a similar way, C and D can be said to have the same cross-mix scale size

$$S(C) = S(D) \Rightarrow S(A) > S(D). \qquad (6\text{-}15)$$

However, from proposition 2, D has a larger cross-mix scale size than A. Therefore, if the isoquants of the frontier used to aggregate the inputs and outputs do not cross, we have a unique relative cross-mix scale size in multiple dimensions.

In order to obtain a CRS function that has non-crossing isoquants, homotheticity must be imposed on the CRS function. This means that each isoquant in input space is uniquely defined by an isoquant in output space and is a multiple of the isoquant for the unit output vector.

Once the inputs and outputs have been aggregated using an homothetic CRS frontier, we can give each unit a unique cross-mix scale size relative to a reference unit.

Note that other problems may occur with a non-homothetic technology in multiple dimensions. Schmidt (1985) points out a problem with the measurement of allocative efficiency for non-homothetic technologies; "Another potential problem ... can occur if technology is not homothetic, so that the cost-minimizing input proportions depend on output." Note that this is not a problem in the single output case, as the CRS function is always homothetic but it will become a problem in multiple dimensions unless account is taken of the 'cross-mix' efficiency which will be defined in the next section.

6.5.1 Imposing homotheticity on the SF translog function

In the single-input or output case, any CRS function is homothetic (see Shephard (1970)). However, in multiple dimensions, homotheticity must be imposed on the CRS function ("Homogeneous correspondences are not necessarily homothetic", Färe and Shephard (1976)). In the SF method this is possible by restricting the form of the estimating function. In order to show how this is possible the translog function is taken as the SF estimating function and the restrictions of CRS, and then homotheticity, will be imposed.

6.5.1.1 Imposing CRS on the translog function: the single-output case

A function f, mapping multiple inputs, $\mathbf{x} \in \Re^n$, to a single output $y = f(\mathbf{x})$, has CRS if

$$y = f(x) \Leftrightarrow \mu y = f(\mu x), \forall\ \mu \in \Re^{+}. \qquad (6\text{-}16)$$

The general form of the translog function in the single output case is

$$\ln(y) = \ln(A) + \sum_{i=1}^{m} \alpha_i \ln(x_i) + \sum_{j=1}^{m}\sum_{k=1}^{m} \beta_{jk} \ln(x_j)\ln(x_k) \qquad (6\text{-}17)$$

where $\mathbf{x} \in \Re^m$ is the vector of inputs (Christiansen, Jorgenson and Lau (1971)).

To impose CRS on this function, the parameters must be restricted so that:

- $\sum_{i=1}^{m} \alpha_i = 1$,
- $\sum_{j=1}^{m}\sum_{k=1}^{m} \beta_{jk} = 0$ and
- $2\beta_{jj} + \sum_{k=1}^{m}\sum_{l=1}^{m} \beta_{kl} = 0$ for $k \neq l$.

(Christiansen, Jorgenson and Lau (1971)).

For example, in the case of a single output and two inputs, the translog function is given by

$$\ln(y) = \ln(A) + \alpha_1\ln(x_1) + \alpha_2\ln(x_2) + \beta_{11}(\ln(x_1))^2 + \beta_{22}(\ln(x_2))^2 + (\beta_{12} + \beta_{21})\ln(x_1)\ln(x_2) \quad (6\text{-}18)$$

and the CRS restrictions are

$$\alpha_1 + \alpha_2 = 1,\ \beta_{11} + \beta_{12} + \beta_{21} + \beta_{22} = 0,$$

$$2\beta_{11} + \beta_{12} + \beta_{21} = 0 \text{ and } 2\beta_{22} + \beta_{12} + \beta_{21} = 0.$$

6.5.1.2 Imposing CRS on the translog function: the multiple-output case

In the case of multiple inputs and outputs, the function $r = g(\mathbf{x},\theta)$ has CRS if

$$r = g(x,\theta) \Leftrightarrow \mu r = g(\mu x,\theta),\ \forall\ \mu \in \Re^+. \quad (6\text{-}19)$$

That is, if all the inputs are increased by the factor μ and the output mix remains constant, then the length of the output vector, r, must also be increased by the same factor.

In the multiple output case, the general form of the translog function is

$$\ln(r) = \ln(A) + \sum_{i=1}^{m} \alpha_i \ln(x_i) + \sum_{j=1}^{s-1} \beta_j \ln(\theta_j) + \sum_{k=1}^{m} \sum_{l=1}^{m} \gamma_{kl} \ln(x_k) \ln(x_l) + \sum_{p=1}^{s-1} \sum_{q=1}^{m} \delta_{pq} \ln(\theta_p) \ln(x_q) + \sum_{o=1}^{s-1} \sum_{n=1}^{s-1} \lambda_{no} \ln(\theta_o) \ln(\theta_n) \quad (6\text{-}20)$$

where $\mathbf{y} = r \cdot m(\theta) \in \Re^s$ and $\mathbf{x} \in \Re^m$.

Theorem 1

To impose CRS on this function, the restrictions on the parameters are

- $\sum_{i=1}^{m} \alpha_i = 1,$
- $\sum_{k=1}^{m} \sum_{l=1}^{m} \gamma_{kl} = 0,$
- $2\gamma_{jj} + \sum_{k=1}^{m} \sum_{l=1}^{m} \gamma_{kl} = 0$ for $k \neq l$ and
- $\sum_{q=1}^{m} \delta_{pq} = 0$ for $p = 1,\ldots,s\text{-}1.$

For the proof, see Appendix 4.

Now, in the single output case, by imposing CRS on the frontier, homotheticity has also been imposed. However, in the multiple input and output case homotheticity still needs to be imposed.

6.5.1.3 Imposing homotheticity on the CRS translog function

Homotheticity can be defined on the input set or the output set or both the input and output sets. A frontier is said to be **input homothetic** (Hanoch (1970)) if each input isoquant is a multiple of the unit isoquant in input space (i.e. the isoquant which corresponds to the unit output vector). Similarly, a frontier is said to be **output homothetic** if each output isoquant is a multiple of the unit isoquant in output space. If the frontier is input homothetic *and* output homothetic then the frontier is said to be **inversely homothetic** (Färe and Primont (1995)).

If a function is homogeneous of degree 1 (i.e., CRS) and separable (all homothetic functions must be separable (Hanoch (1970), Shephard (1970))), then it must be inversely homothetic (Hanoch (1970)). Therefore the term homothetic CRS will be used to mean a frontier which has CRS and is inversely homothetic.

A CRS frontier is homothetic if each input isoquant uniquely defines an output isoquant (i.e., any point on the input isoquant can produce any point on the output isoquant and vice versa.) The input isoquant is a function of the two inputs $g(x_1,x_2)$. As each isoquant is a radial projection of every other isoquant, $g(x_1,x_2) = c_1$ and c_1 increases as the isoquant moves further from the origin. Similarly, the output isoquant is a function of the two outputs, or equivalently, of r and θ, giving $f(r,\theta) = c_2$. As the frontier has CRS, as c_1 increases by a factor μ, c_2 must

increase by the same factor μ. Therefore, the full transformation, or production, function is given by $f(r,\theta) = \rho \cdot g(x_1,x_2)$, i.e. the function must be input-output separable. This is stated by Shephard (1970) p.273 as Corollary F: "The distance function D(x,y) takes the separable form f(x)/g(y) if and only if the input structure of the production correspondence is homothetic."

In the case of the CRS translog function

$$\ln(r) = \ln(A) + \sum_{i=1}^{m} \alpha_i \ln(x_i) + \sum_{j=1}^{s-1} \beta_j \ln(\theta_j) + \sum_{k=1}^{m}\sum_{l=1}^{m} \gamma_{kl} \ln(x_k)\ln(x_l)$$
$$+ \sum_{p=1}^{s-1}\sum_{q=1}^{m} \delta_{pq} \ln(\theta_p)\ln(x_q) + \sum_{o=1}^{s-1}\sum_{n=1}^{s-1} \lambda_{no} \ln(\theta_o)\ln(\theta_n)$$

$\sum_{i=1}^{m} \alpha_i = 1$, $\sum_{k=1}^{m}\sum_{l=1}^{m} \gamma_{kl} = 0$, $2\gamma_{jj} + \sum_{k=1}^{m}\sum_{l=1}^{m} \gamma_{kl} = 0$ for $k \neq l$ and $\sum_{q=1}^{m} \delta_{pq} = 0$, $p = 1,\ldots,s-1$, the extra restriction to impose homotheticity is $\delta_{pq} = 0 \; \forall \; p = 1,\ldots,r-1$ and $q = 1,\ldots,m$.

So the two-input, two-output homothetic CRS translog function is given by

$$\ln(r) = \ln(A) + \alpha\ln(x_1) + (1-\alpha)\ln(x_2) + \beta\ln(\theta)$$
$$+ \gamma(\ln(x_1))^2 + \gamma(\ln(x_2))^2 - 2\gamma\ln(x_1)\ln(x_2) + \lambda(\ln(\theta))^2. \qquad (6\text{-}21)$$

6.5.2 Imposing homotheticity in DEA

In order to impose homotheticity in DEA, a new model is required. This model must involve fewer constraints or more variables than the Charnes, Cooper and Rhodes model, as the homothetic CRS frontier envelops the data less closely than the CRS frontier. The development of this model is left as future research.

This model will measure how much more a unit could increase its output by, for any mix of inputs equivalent to its current levels. It is proposed here that this model will measure the full technical efficiency. The ratio of the efficiency measure estimated by the new model, to that estimated by Model 1, the CRS model, will give a measure of the **cross-mix efficiency** of the unit, i.e. the extra output that could be achieved once CRS efficient if the DMU changes its mix of inputs. This cross-mix efficiency is only apparent in the case of at least two inputs and two outputs.

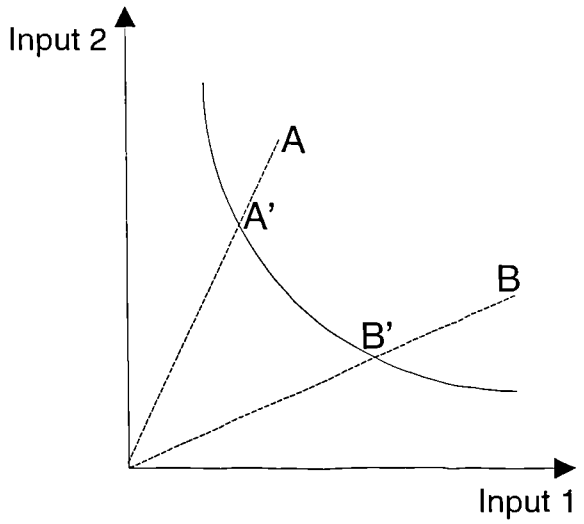
6.5.3 Using the Malmquist index to measure cross-mix scale size in multiple dimensions

Using the distance function in the same way as in Chapter 3, it is possible to compare units across different input (output) mixes. Note however, that the distance function must now be homothetic CRS.

Take two units, A with inputs $x_A=(x_{1a}, x_{2a})$ and outputs $y_A=(y_{1a}, y_{2a})$ and B with inputs $x_B=(x_{1b}, x_{2b})$ and outputs $y_B=(y_{1b}, y_{2b})$ so that A and B do not have the same input or output mix.

Consider the two units in input space.

Figure 6-3. Cross-mix scale size in multiple dimensions



The isoquant shown is that of (y_{1a}, y_{2a}), the output vector of unit A. Both A and B can be projected onto this isoquant at points A' and B' respectively. These are the efficient points with the same input mixes as A and B and the output levels of A. The isoquant for the output levels of unit B could have been chosen just as well, as the distance function is homothetic CRS so the isoquant for the output vector of B must be a radial expansion or contraction of the isoquant for the output vector of A.

Now the size of B relative to the size of A is given by;

$$\frac{S(B)}{S(A)} = \frac{S(B)}{S(B')}\frac{S(B')}{S(A')}\frac{S(A')}{S(A)}$$
$$= \frac{S(B)}{S(B')}\frac{S(A')}{S(A)} \qquad (6\text{-}22)$$
$$= \frac{D_I^{CRS}(x_B, y_A)}{D_I^{CRS}(x_A, y_A)}$$

Once again, this is the Malmquist input quantity index.

For example, if this relative cross-mix scale size gives a value of 1.5, B can be said to be operating at a larger scale size than A. This actually means that if A could change its input mix so that it had the same mix as B, keeping its output levels constant, these equivalent inputs levels would be two thirds of those of B in a radial direction.

6.6 Operationalising the cmss measure

In this section a cross-mix scale size measure will be defined on a data set with two inputs and two outputs, generated according to DGP D in Appendix 2. This data set is generated from an input-output separable Cobb-Douglas function

$$y_1^{0.4} y_2^{0.6} = A\, x_1^{0.5} x_2^{0.6} \qquad (6\text{-}23)$$

The SF method can be used to estimate the cross-mix scale size.

Firstly, CRS homotheticity must be imposed on the estimating function. This is simple if a Cobb-Douglas function is used, as the CRS function is immediately homothetic. In the case of the translog function the restrictions must be imposed as in (6-19). The stochastic frontier can then be estimated from the data. One unit must be chosen as the reference unit. The cross-mix scale size can then be computed using the following five-step procedure:

Step 1: Using the observed input vectors $\mathbf{x}_j$ and the corresponding output vectors y_j of DMUs j = 1,...,N, where N is the number of DMUs and $\mathbf{x}_j = (x_{j1}, x_{j2},...,x_{jm})$, estimate a production function $r = f(\theta,\mathbf{x}; \beta)$ with restrictions to impose homothetic CRS.

Step 2: Select an arbitrary observed output vector, $(r,\theta)^r$, to give the **reference** isoquant in input space. This isoquant is defined to have a cross-mix scale size of 1.

Step 3: To compute the cross-mix scale size of some DMU j, compute first the ratios

$$r_{ji} = \frac{x_{ji}}{x_{j1}}, \quad i = 2,...,m. \qquad (6\text{-}24)$$

(If $x_{j1} = 0$ use another input which is non-zero as a 'reference input'.) This gives each input level x_{ji}, i = 2,...,m of DMU j in terms of x_{j1}:

$$x_{ji} = r_{ji}\, x_{j1} \qquad (6\text{-}25)$$

Step 4: Substitute for x_{ji}, i = 2,...,m from (6-25) into $g(r,\theta)^r = f(\mathbf{x})$ and solve for x_{j1}. Let the resulting value be x^r_{j1}. This is the level of input 1 when the input vector $\mathbf{x}_j$ of DMU j is expanded or contracted to the reference isoquant.

Step 5: The cross-mix scale size of DMU j is x_{j1}/x^r_{j1}.

Any one of the inputs could have been used to compute the cross-mix scale size and not just input 1, as $x_{j1}/x^r_{j1} = x_{ji}/x^r_{ji}\ \forall\ i = 2,..,m$.

The cross-mix scale sizes for the data set from DGP D have been estimated using this five-step procedure.

Figure 6-4. Scale efficiency across scale size

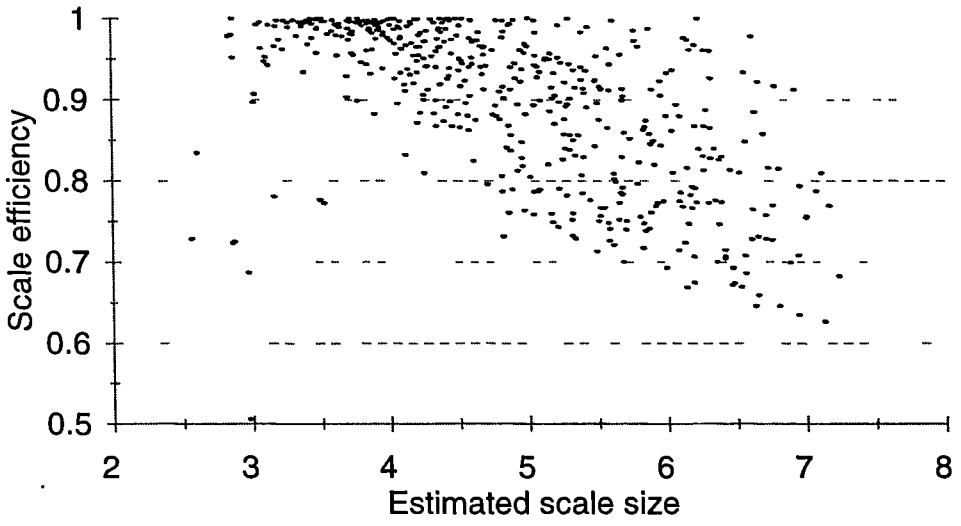


Figure 6-4 shows the scale efficiency estimates under DEA plotted against the estimated cross-mix scale size. There obviously is a pattern (although it is much less clear than in the single output case): The scale inefficiency is increasing as the estimated scale size increases.

6.7 Conclusions

It has been shown in this chapter that it is possible to define a multiple-output, multiple-input production frontier in both DEA and SF methods that can be used to define a relative cross-mix scale size measure.

In order to do this the SF method has been used (In the DEA method a new model must be developed to impose homotheticity on the PPS).

A simple two-input, two-output example was given to show how it is possible to calculate the relative cross-mix scale sizes.

Now that scale issues in DEA and SF methods have been covered for all sets of data, the next chapter will consider differences between the methods across input mix.

Chapter 7

Functional misspecification in the SF method

leading to variation of fit across input mix

7.1 Introduction

Now that possible differences across scale size have been investigated, this chapter will investigate variation of fit across input mix.

In this chapter, two hypotheses will be examined; Hypotheses 5 and 6. Hypothesis 5 proposes that if there are few DMUs at extreme input mixes then the DEA efficiency estimates for units at these mixes will be given overestimates of the true efficiencies. Hypothesis 6 is related to functional misspecification in the SF method and has already been illustrated for misspecification across scale size in Chapter 5. In this chapter it is shown how misspecifying the form of the production function in the SF method can lead to functional misspecification across the input mix. Hypothesis 6 states that if there is functional misspecification in the SF method there will be regions of the technology where the SF efficiency estimates are less then the true values and regions where they are greater than the true values. Here the case where the DMUs with extreme input mixes (DMUs with a very high or very low ratio of inputs) are in areas of poor specification by the SF estimating function will be investigated. It is shown in Section 7.4 that it may be possible to identify this misspecification by comparing the SF and DEA estimates.

The functional misspecification in the example shown here is caused by the estimating function imposing a very different elasticity of substitution

to that of the underlying technology. In this case the underlying technology has a very low elasticity of substitution (0.3) and a Cobb-Douglas function is used as the estimating function (which has an elasticity of substitution of unity).

A two-input, single-output simulated technology will be used, as described in Data Generating Process C in Appendix 2. The effect of the form of the technology imposed by the SF method will be investigated when there is no random noise in the data. In particular, the effects on the estimates of the SF method will be investigated when the underlying function is not well estimated by the estimating function to see whether it is possible to identify this functional misspecification by comparing the SF and DEA estimates.

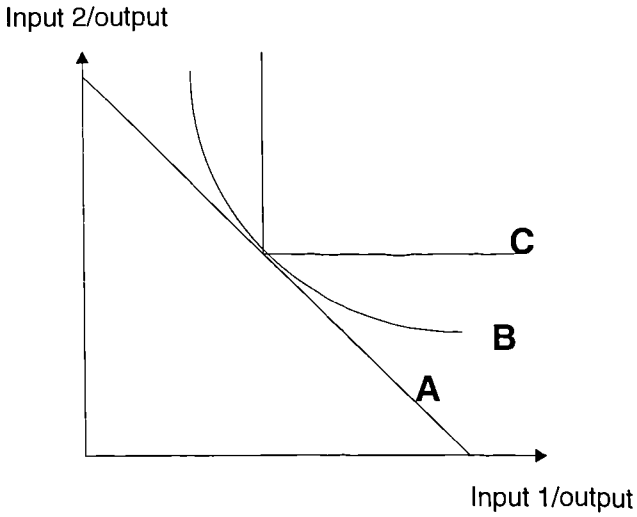
The next section will show how variation of fit can occur across input mix. Section 7.3 will give the results for a simple example to illustrate Hypotheses 5 and 6. Section 7.4 will explain how the results can be used to gain an insight into the nature of the underlying data and Section 7.5 will present the conclusions of this chapter.

7.2 Variation of fit across input mix

Variation of fit is illustrated here using the Cobb-Douglas function as the SF estimating function. In this case the DMUs with extreme input mixes are likely to be misclassified. Consider the isoquant of a production

function under constant returns to scale. The isoquant is the locus of input levels needed to efficiently produce one unit of output. (See Figure 7-1, curve B.) The curvature of the isoquant depends on the elasticity of substitution between the inputs. An elasticity of substitution of zero would be given by a function with a rectangular isoquant (see C in Figure 7-1) and a value of infinity would be given by a function with a straight line isoquant (see line A in Figure 7-1). The Cobb-Douglas function (line B in Figure 7-1) has an elasticity of 1. Functions with other elasticities of substitution will fit in between A and C, Figure 7-1. If A or C is the underlying technology, and B the estimating function, DMUs operating under extreme input mixes (high ratio of one input to another) will be missclassified on efficiency, while DMUs operating under other input mixes will have their efficiencies estimated well.

Figure 7-1. Elasticities of substitution



Variation of fit in this case will be across input mix. What would be expected in this case - which of SF or DEA will do better? Recall the hypotheses constructed in Chapter 2 which are relevant to variation of fit across input mix: Hypothesis 5 indicates what would be expected in the case of DEA;

Hypothesis 5

If the technology has few units in certain regions of input or output mix, and these regions are at the edge of the technology, then the DMUs in these regions may be given estimates such that $E_{DEA} > E_{TRUE}$.

and Hypothesis 6, the case of SF;

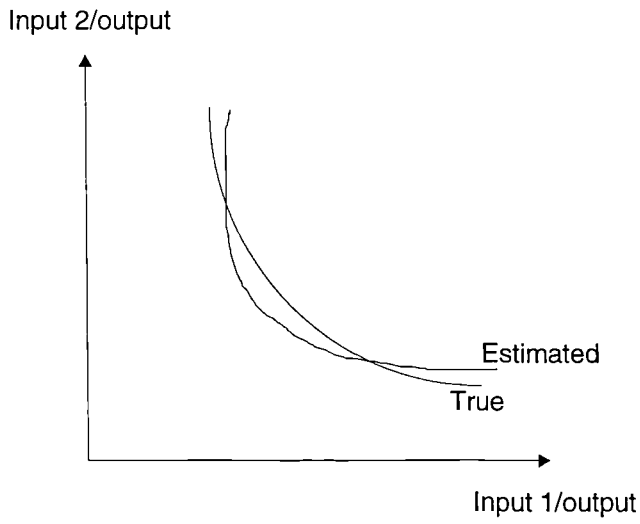
Hypothesis 6

If the true frontier is not well estimated by the SF function, then the estimated efficiencies will have regions where they are greater than and less than the true efficiencies across scale size or input mix, depending on whether the misspecification varies across scale size or input mix

In the example which will be examined here, the variation of fit is expected to occur across input mix as the underlying function has a low elasticity of substitution and the Cobb-Douglas function has elasticity of substitution fixed at 1.

We would expect the relation between the estimated and true functions to be similar to that shown in Figure 7-2.

Figure 7-2. The estimated SF function



7.3 The results

7.3.1 Illustrating variation of Fit across input mix

In order to investigate functional misspecification in the SF method a Cobb-Douglas function is used as the estimating function.

The Cobb-Douglas function is given by;

$$y_{obs} = A\, x_1^{\alpha}\, x_2^{\beta}\, e^{s-t} \qquad (7\text{-}1)$$

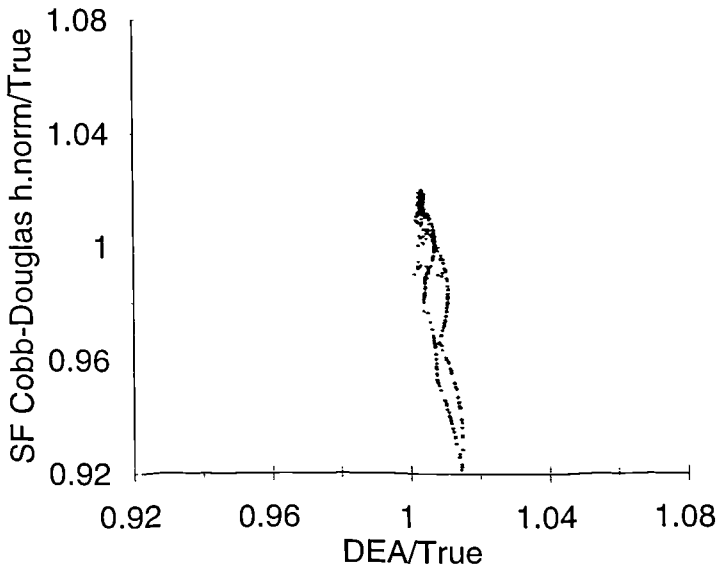
where $e^{s} = v$, $e^{-t} = 1 - u$, and y_{obs} is the observed output, x_1, x_2 the two inputs, v the random noise and u the inefficiency.

Taking logs gives a linear function:

$$Y = a + \alpha X_1 + \beta X_2 + s - t \qquad (7\text{-}2)$$

where $Y = \ln y_{obs}$, $X = \ln x$, $a = \ln A$. The restriction of constant returns is given by: $\alpha + \beta = 1$.

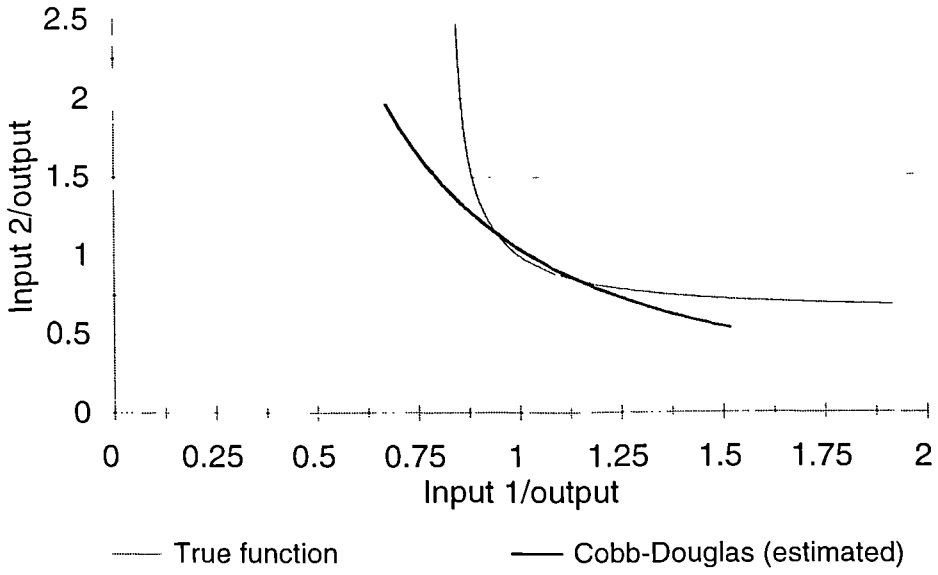
Figure 7-3. The DEA results compared to SF Cobb-Douglas



The ratios of the SF Cobb-Douglas efficiency estimates to the true values are plotted against the ratios of the DEA estimates to the true efficiency values in Figure 7-3[1].

From this graph it is clear that the results from the SF method are much worse than those from the DEA method.

Figure 7-4. Isoquant of the underlying function in comparison with the Cobb-Douglas isoquant (no random noise)



In Figure 7-4, the estimated and underlying isoquants are plotted. There are three clear regions across input mix. Firstly, when the ratio

[1] To ensure that no other assumptions are violated, the random noise term was assumed to be normally distributed and the inefficiency is assumed to be distributed with a half-normal distribution. As the underlying technology has been generated with CRS, CRS has been imposed in both the DEA and SF methods.

of input 1 to input 2 is large, the estimated function lies below the true function so the estimated efficiencies will be much less than the true efficiencies. As the ratio of input 1 to input 2 decreases, the estimated frontier again lies above the true frontier so the estimated efficiencies will again be less than their true values. In the central input mix range, around the value of the ratio of input 1 to input 2 equal to 1, the estimated function lies slightly above the true function. The DMUs in this range of input mixes will be given estimates which are slightly greater than their true values.

Table 7-1. Mean absolute deviations and mean deviations

Estimating method	MAD	Mean deviation
DEA	0.00700	-0.00700
Cobb-Douglas SF	0.03099	0.02346
Translog SF	0.00287	-0.00084

The mean deviation and mean absolute deviations of the estimated to the true efficiency values are given in Table 7-1.

From this table it is clear that the Cobb-Douglas SF method is giving much worse estimates than either DEA or the translog SF method. As Hypothesis 6 states, it has been demonstrated that, when a restrictive assumption (i.e. in this case that the elasticity of substitution is equal to

1) is imposed on the SF method, the SF method gives estimates which are greater than the true values for some units and less than the true values for other units. This difference is shown in the table as the smaller mean deviation for the Cobb-Douglas method as opposed to the higher mean absolute deviation.[2] The fact that these differences are in specific regions of the technology (i.e. different input mixes) will be demonstrated more clearly in the next section by examining the nature of the functional deviation across the input mix.

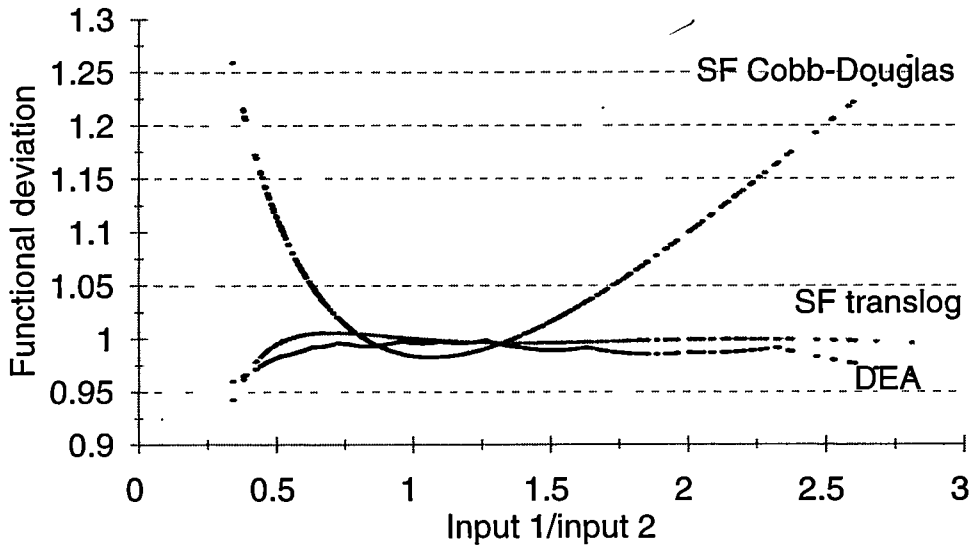
7.3.2 Functional deviation

In order to be able to compare how the methods performed on the individual DMUs, the Functional Deviation value ((estimated efficient output)/(true efficient output)) for each DMU was calculated under each method of estimation. Because the Cobb-Douglas function has been constrained to have constant returns to scale, the functional deviation will only occur across input mix.

To compare how DEA and SF perform on the DMUs for which the Cobb-Douglas function is badly estimating the underlying technology, the mean absolute deviation for just the DMUs for which Cobb-Douglas had a FD greater than 1.15 were calculated. The summary statistics

[2] It is clear from this table that there is also some functional deviation in the translog SF method, i.e. some of the estimates are above the frontier and some are below but this difference is much smaller.

Figure 7-5. Functional deviation for each method across input mix



are given in Table 7-2 for comparison with the results when considering the whole set of data.

These results show that DEA performs much better than the Cobb-Douglas SF method in areas where the Cobb-Douglas function misspecifies the underlying technology (i.e. units at the extreme input mixes). Note that the DEA results at the extreme mixes are, however, given slightly worse estimates than the average. This is also clearly shown by the graph in Figure 7-5.

Clearly in DEA the mix has little impact on deviations except at the very extreme input mixes where there are fewer units and the DEA method is overestimating the true efficiencies as expected from Hypothesis 5.

Table 7-2. Mean absolute deviations for the DMUs which have FD values above 1.15 under half-normal Cobb-Douglas SF in comparison with the MAD values for all the DMUs

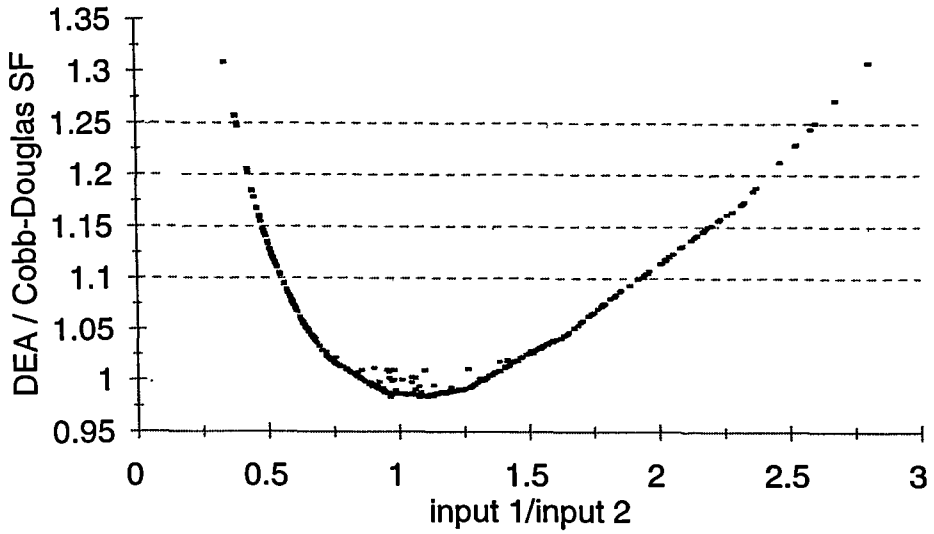
	MAD for extreme DMUs	MAD for all DMUs
DEA	0.02015	0.00700
SF Cobb-Douglas	0.13679	0.03099
SF Translog	0.01161	0.00287

This section has shown that the two hypotheses, 5 and 6, have been illustrated in this simple example. How can this information be used to gain information about the underlying technology when it is unknown?

7.4 Using the results to identify the true nature of the underlying technology

In Figure 7-6, the ratio of the DEA estimates to the SF Cobb-Douglas efficiency estimates is plotted against the input mix. This is the graph which could be plotted from the actual data without having any knowledge of the underlying technology. There is clearly a difference between the estimated frontiers of DEA and SF. As the results follow such a precise relationship it is also clear that the random noise levels must be very low. The DEA estimates are almost all greater than the SF estimates.

Figure 7-6. Ratios of the DEA estimates to the SF estimates across input mix (no random noise)



For the DMUs which have $E_{DEA} > E_{SF}$ we have five possibilities for the DMUs:

1. $E_{DEA} > E_{SF} > E_{TRUE}$
2. $E_{DEA} > E_{SF} = E_{TRUE}$
3. $E_{DEA} > E_{TRUE} > E_{SF}$
4. $E_{DEA} = E_{TRUE} > E_{SF}$
5. $E_{TRUE} > E_{DEA} > E_{SF}$

For the DEA efficiencies to be greater than the true efficiencies, the frontier must be pulled towards the units for some reason. Across input mix this could only be because there are not enough efficient units at the extreme input mixes. Looking at the number of units which have large deviations between the DEA and SF estimates this does not seem

to be the likely cause. This makes the first two possibilities unlikely. The final three possibilities all have $E_{SF} < E_{TRUE}$. For the SF estimates to be less than the true values, and the deviations varying across input mix, it is likely that there is functional misspecification in the SF model and this is due to some sort of restriction on the elasticity of substitution of the estimating function. In the case of the Cobb-Douglas function this is clearly a possibility and should lead to the analyst trying a more flexible form in the SF method.

7.5 Conclusions

This chapter has investigated the ways in which SF and DEA can misspecify the true frontier for certain ranges of input mix. In the case of the SF method, it has been shown that, by imposing a restriction on the elasticity of substitution of the production function, the estimated function does have regions where it lies above the true function and regions where it lies below the true function, illustrating Hypothesis 6.

Similarly, in the case of DEA, it was found that the fewer units at the extreme input mixes led to the DEA efficiency estimates being greater than the true values for these units, illustrating Hypothesis 5.

In the case which has been considered, the DMUs which were operating under areas of the technology which were badly misspecified were those with extreme input mixes. It is valuable to know which

method is likely to give better estimates for these DMUs, and under which circumstances, because these DMUs may

- have pioneering operating methods,
- have unusual value systems which should not be penalised,
- be dealing with unusual environments.

The approach outlined in this chapter could be used to identify cases where there might be variation of fit across input mixes. For example, we could estimate the DEA efficiencies and we would expect these to be invariant with input mix (except for the very extreme input mixes). Therefore, plotting a comparison of the DEA estimates to those of an SF method across input mix should identify whether there is variation of fit across input mix in the SF case.

The next chapter will summarise the use of the hypotheses which were outlined in Chapter 2, in order to gain extra insight into the underlying technology and to give indications as to which of DEA or SF methods may be giving the better estimates.

Chapter 8

An algorithm for applying the results

8.1 Introduction

The previous chapters have investigated the ways in which DEA and SF methods can give inaccurate estimates of the true efficiency values for units on scale size, input mix or across the whole technology.

In this chapter, the ways in which the conclusions about the hypotheses outlined in Chapter 2 can be exploited in a real data set in order to arrive at more accurate efficiency estimates will be discussed. It will be shown how a comparison between the results of DEA and SF can give an indication as to which of the methods is giving the most accurate estimates of efficiency in different regions of the technology.

In order to draw conclusions about the performance of the methods, these steps need to be followed:

1. Identify all the assumptions which the methods make[1].
2. Identify the regions where the methods give different results. Are these regions on input mix or scale size or across the whole technology?
3. Try to identify which assumption(s) may not be holding by drawing on the hypotheses in Chapter 2.
4. If possible test these assumptions.

[1] If more assumptions are identified than were outlined in Chapter 2, then the conclusions given in Table 8-1 of this chapter will change.

5. Decide which of the methods is giving the closer estimates in each region of the technology.

The difficult step is 3, that is deciding which of the assumptions may not be holding. The next section will discuss this problem and summarise the results from the previous chapters.

8.2 Identifying which of the assumptions may not be holding

For each DMU, the estimate from one method can only be greater than, less than or equal to the estimate from the other method. These differences between the estimates may be related to the input mix or the scale size, or they may vary randomly across the whole technology[2].

Once the regions have been identified where the methods differ, how is it possible to tell which of the assumptions of the methods may not be holding? In order to answer this question, each of the assumptions of each method needs to be considered in turn, as was done in Chapter 2. Then, depending on the nature of the differences between the estimates of the methods, it is possible to draw on the hypotheses

[2] The differences may also vary across both the input mix and the scale size but this possibility is not considered here as we only allow for a single assumption being violated in each method.

illustrated in the previous chapters to identify which of the assumptions may not be holding.

In each of the following sections, note that it is assumed that only one assumption is violated in each method at a time. That is, if there is random noise in the data, it is assumed that there is no functional misspecification in the SF method, there are enough efficient observations well spread out in the DEA method, etc.

8.2.1 Differences across the whole technology

Differences between the estimates which occur in the same fashion throughout the technology were examined in Chapter 3.

The conclusions found in that chapter were:

- If all assumptions of both methods are met, the DEA estimates will still involve a finite sample error leading to $E_{DEA} > E_{TRUE}$. However, as the sample size increases, this error decreases and the estimates are approximately equal to the true values. (Throughout this chapter $E_{DEA} \cong E_{TRUE}$ will be used to denote cases where no assumptions are violated in DEA.)
- If there is random noise in the data, the estimates will be such that $\overline{E}_{TRUE} > \overline{E}_{SF} > \overline{E}_{DEA}$ unless there are very low levels of random noise and the SF method identifies the average magnitude of this noise leading to the inequality $\overline{E}_{TRUE} \cong \overline{E}_{SF} > \overline{E}_{DEA}$.

- If the SF method assumes an incorrect distribution for the inefficiency term, the SF method will underestimate all the inefficiencies leading to the inequality $E_{TRUE} \cong E'_{DEA} > E_{SF}$

These are the only possibilities that have been identified when assumptions are violated across the whole technology.

Therefore, if across the whole technology it is found that

- $\overline{E}_{SF} > \overline{E}_{DEA}$, it is possible to conclude that the data contains random noise and the estimates from both methods are, on average, less than the true efficiency values.
- $E_{DEA} > E_{SF}$ across the whole technology, a comparison of the efficiency estimates across the DEA estimates should be performed (see Figure 5-7) to identify possible misspecification of the inefficiency distribution in the SF method.

In the next two sections, the possible differences between the true efficiencies and the estimates will be identified for assumptions which affect the methods across the scale size and then the input or output mix. Throughout these sections, the nature of the estimates to the true efficiencies will be highlighted. The findings will then be summarised in Table 8-1 for the observed differences between the results from the two methods when the true efficiency values are unknown. Examples of how this table can be used will be given in Sections 8.3.2 and 8.3.1.

8.2.2 Differences which vary across scale size

Differences between the estimates across scale size were identified in Chapter 5.

If there are differences between the estimates across scale size, then a graph plotting the ratio of the estimates across the estimated scale size should identify regions where the estimates are closer and regions where they are further apart (see Figure 4-18).

The reasons for differences between the estimates across scale size must lie in the assumptions which the methods are making about the nature of the returns to scale of the technology or the convexity of the PPS.

If the SF method imposes CRS on the technology unnecessarily, then the SF efficiency estimates will be greater than the true values in some scale size ranges and less than the true values in others (see Figure 5-5). If this is the only assumption violated, then $E_{SF} > E_{DEA} \cong E_{TRUE}$ in some ranges of scale size and $E_{DEA} \cong E_{TRUE} > E_{SF}$ in others.

If the DEA method imposes an unnecessarily strict restriction (e.g. CRS) on the nature of the returns to scale in a certain region of the technology, then DEA estimates of the efficiencies of the units in this region will be given underestimates of the true values. In the region where the restriction does not hold $E_{SF} \cong E_{TRUE} > E_{DEA}$ (see Figure 5-3).

On the other hand, if the DEA method does not impose a necessary restriction on the technology then the regions where there are not enough efficient units to properly define the frontier will lead to estimates such that $E_{DEA} > E_{SF} \cong E_{TRUE}$ (see Figure 5-4).

If both methods impose a too restrictive RTS assumption, then the SF restriction will lead to ranges of scale size where $E_{SF} > E_{TRUE}$ and other ranges where $E_{TRUE} > E_{SF}$. The restriction on the DEA method will give certain ranges of scale size where $E_{TRUE} > E_{DEA}$ and others where $E_{DEA} \cong E_{TRUE}$. How large each of these ranges is and how they coincide depends on the true nature of the RTS. This gives the possible inequalities:

- $E_{SF} > E_{TRUE} > E_{DEA}$ in regions where the DEA RTS assumption is too strict and the SF restriction leads to the SF frontier lying below the true frontier in these regions;
- $E_{SF} > E_{TRUE} \cong E_{DEA}$ in regions where the DEA RTS assumption is correct but the SF restriction still leads to the SF frontier lying below the true frontier in these regions;
- $E_{TRUE} > E_{SF} > E_{DEA}$, where the DEA RTS assumption is too strict and the SF estimated frontier lies above the true frontier;
- $E_{TRUE} > E_{DEA} > E_{SF}$ as above, but the SF frontier lies above the DEA frontier as well as the true frontier;
- $E_{TRUE} \cong E_{DEA} > E_{SF}$. In this case the DEA RTS assumption is correct, but the SF frontier lies above the true frontier;

- $E_{TRUE} > E_{SF} \cong E_{DEA}$ both frontiers lie above the true frontier due to the RTS restriction.

These inequalities will only hold in certain ranges of scale size and these ranges will depend on how large the region is where the restrictive assumption does not hold.

These possibilities are summarised in Table 8-1, which shows how the findings are of use in a real application where the true efficiencies are unknown.

It is also possible that a too restrictive assumption about RTS is imposed on the SF method by using a restrictive estimating function (e.g. the Cobb-Douglas function), while a more restrictive assumption about the returns to scale is needed in the DEA method but is not used. Once again the scale size will be split into ranges where $E_{SF} > E_{TRUE}$ and $E_{SF} < E_{TRUE}$. However, the DEA estimates will now be $E_{DEA} > E_{TRUE}$ in some ranges of scale size and $E_{DEA} \cong E_{TRUE}$ in other ranges. This will lead to possible inequalities of the form

- $E_{SF} > E_{DEA} > E_{TRUE}$ in regions where there are not enough efficient units for the DEA method to identify the true nature of the returns to scale and the SF frontier lies below the true frontier;
- $E_{DEA} > E_{SF} > E_{TRUE}$ as above but the SF frontier now lies above the true frontier and the DEA frontier;

- $E_{DEA} \cong E_{TRUE} > E_{SF}$ - the DEA method correctly identifies the nature of the returns to scale and the SF frontier lies above the true frontier;
- $E_{SF} > E_{DEA} \cong E_{TRUE}$ - the DEA method correctly identifies the nature of the returns to scale and the SF frontier lies below the true frontier;
- $E_{DEA} > E_{TRUE} > E_{SF}$ - there are not enough efficient units for the DEA method to identify the true frontier and the SF frontier lies above the true frontier.

If the PPS is non-convex then the DEA efficiency estimates will be less than the true values in the region of non-convexity. How the SF estimates relate to the true values will depend on whether the SF method imposes convexity or not. If the SF method allows for a non-concave frontier then the region of the PPS which is non-convex will have estimates such that $E_{SF} \cong E_{TRUE} > E_{DEA}$ (Hypothesis 7). If the SF method does not allow for a non-concave frontier (i.e. non-convex PPS), then once again there will be variation of fit of the SF frontier to the true frontier leading to certain ranges of scale size where $E_{SF} > E_{TRUE}$ and others where $E_{TRUE} > E_{SF}$. So the underlying inequalities could include:

- $E_{TRUE} > E_{SF} \cong E_{DEA}$ - both methods do not allow for a non-concave frontier and are affected in the same way;
- $E_{TRUE} > E_{DEA} > E_{SF}$ - the SF frontier lies above both the DEA frontier and the true frontier due to variation of fit;

- $E_{TRUE} > E_{SF} > E_{DEA}$ - the SF frontier lies between the DEA frontier and the true frontier;
- $E_{SF} > E_{TRUE} > E_{DEA}$ - the SF frontier lies below the true frontier;
- $E_{SF} > E_{TRUE} \cong E_{DEA}$ - the SF frontier lies below the true frontier in a region where the true frontier is concave.

To conclude, if the SF and DEA estimates are compared across scale size, and it is found that there is a pattern to the differences, this could be due to a variety of violations of the underlying assumptions (and this is only allowing for one assumption to be violated at a time in each method). In order to be able to identify which of the assumptions is being violated, the pattern of the differences between the estimates needs to be examined. This will be discussed in Section 8.3.

8.2.3 Differences which vary across (input or output) mix

In Chapter 7, reasons for differences between the estimates from the methods were investigated across input mix for the two input case. By plotting the ratio of the estimates from the two methods across the mix it is possible to see whether there is any variation between the estimates across the mix (Figure 7-6).

It was shown in Chapter 7 that the estimates from a SF assessment can vary across the input mix if the estimating function imposes an incorrect elasticity of substitution on the technology. This is most likely

to be a problem if the Cobb-Douglas function is used as the estimating function. In this case there will be regions of the technology where $E_{SF} > E_{TRUE}$ and regions where $E_{TRUE} > E_{SF}$. How these regions vary will depend on the nature of the true elasticity of substitution.

The assumption which can affect the DEA estimates across the input mix is that there are enough efficient units across the whole technology. If there are not enough efficient units at certain input mixes, the DEA estimates in these regions will be greater than the true values; $E_{DEA} > E_{TRUE}$. At other regions the estimates will be $E_{DEA} \cong E_{TRUE}$.

If both of the above assumptions are invalid, there will be regions of input mix where the estimates are such that

- $E_{DEA} > E_{TRUE} > E_{SF}$ - there are not enough efficient DMUs for DEA to correctly identify the frontier and the SF frontier lies above the true frontier;
- $E_{SF} > E_{DEA} > E_{TRUE}$ - as above, but the SF frontier lies below the true frontier and the DEA frontier;
- $E_{DEA} > E_{SF} > E_{TRUE}$ - as above, but the SF frontier lies between the true frontier and the DEA frontier;
- $E_{SF} \cong E_{DEA} > E_{TRUE}$ - similarly, the SF frontier and the DEA frontier coincide;

- $E_{SF} > E_{DEA} \cong E_{TRUE}$ - in the region where there are enough efficient DMUs for the DEA method, the SF frontier lies below the true frontier;
- $E_{DEA} \cong E_{TRUE} > E_{SF}$ - as above, but the SF frontier lies above the true frontier.

How large these regions are will depend on the nature of the underlying elasticity of substitution and the region of input mix where there are few efficient units.

8.3 An algorithm for using the comparative DEA and SF efficiency estimates to arrive at more accurate estimates

The findings in Section 8.2 above can be generalised into an algorithm for comparing the estimates from the two methods in order to decide which of the two methods is giving the more accurate estimates. The algorithm will be based on Table 8-1 which summarises the results from Section 8.2. This table outlines the possible causes for each of the observed differences across the scale size, input or output mix or the whole technology.

The algorithm (summarised in Figure 8-1):

Step 1: Estimate the efficiencies of the data set using both DEA and SF approaches.

Step 2: Plot the ratio of the estimates across scale size (calculating the scale size as outlined in Chapters 4 and 6).

Is there any pattern to the differences?

If yes: Look in Table 8-1 to identify the possible causes for the differences (see Example 1 which follows).

If no,

Step 3: Plot the ratio of the estimates across each input and output mix.

Is there any pattern to the differences?

If yes: Once again, look in Table 8-1 to identify the possible causes (see Example 2 which follows).

If no,

Step 4: Plot the ratio of the estimates across the level of DEA efficiency.

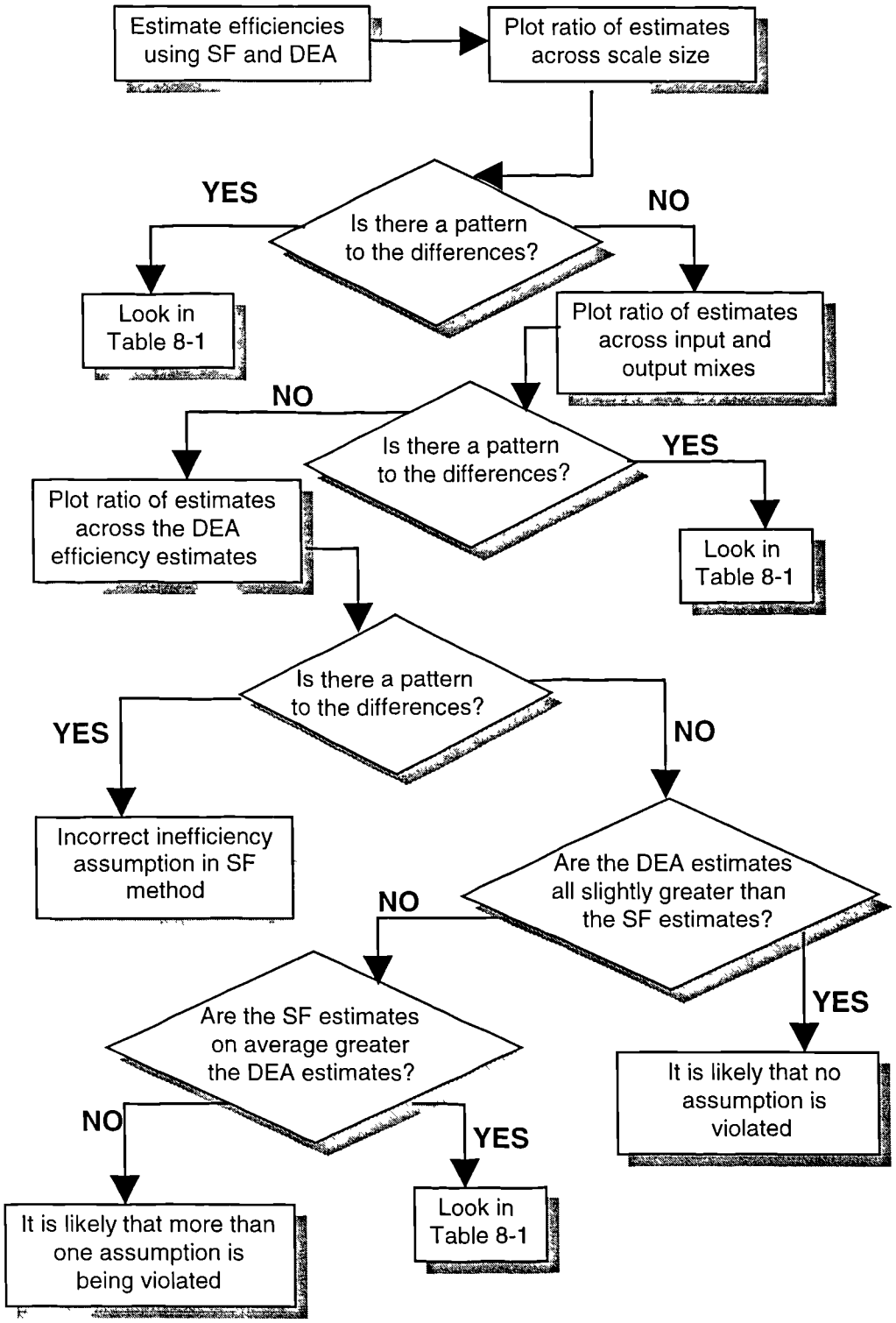
Is there any pattern to the differences?

If yes: It is likely that the assumption about the inefficiency distribution in the SF method is affecting the results. If this is the case, the DEA method is likely to be giving better estimates.

If no,

Step 5: Are all the DEA estimates greater than but close to the SF estimates?

Figure 8-1. An algorithm to identify possible violation of the underlying assumptions of SF or DEA



If yes: No assumption is violated in either method. The DEA estimates will always be slightly higher than the true efficiency values in this case due to finite sample error.

If no,

Step 6: Are the SF estimates on average greater than the DEA estimates?

If yes: It is likely that there is random noise in the data.

If no: The algorithm is not able to identify the causes for the differences between the methods. This may be because more than one assumption is violated.

8.3.1 Example 1

The results of DEA and SF are compared for a particular data set and it is found that the SF estimates are greater than DEA for large scale sizes and less than DEA for small scale sizes.

Look in Table 8-1 at the rows which give possible causes for differences between the efficiency estimates across scale size. The possible causes of the differences at large scale sizes (where $E_{SF} > E_{DEA}$) are;

- too restrictive assumption about RTS in both methods
- too restrictive assumption about RTS in DEA

Table 8-1. Differences between the estimates

Inequality	Whole Tech	mix	scale size	Possible causes	True inequality
$E_{SF} > E_{DEA}$	✓	✗	✗	• random noise – high • random noise – low	$\overline{E}_{TRUE} > \overline{E}_{SF} > \overline{E}_{DEA}$ $\overline{E}_{SF} \cong \overline{E}_{TRUE} > \overline{E}_{DEA}$, or $\overline{E}_{TRUE} > \overline{E}_{SF} > \overline{E}_{DEA}$
	✗	✓*	✗	• not enough efficient DMUs and FD in SF • too restrictive functional form in SF	$E_{SF} > E_{DEA} > E_{TRUE}$ $E_{SF} > E_{DEA} \cong E_{TRUE}$
	✗	✗	✓**	• too restrictive RTS in both methods • too restrictive RTS in DEA • too restrictive RTS in SF • non-convex PPS	$E_{SF} > E_{TRUE} > E_{DEA}$, or $E_{TRUE} > E_{SF} > E_{DEA}$ $E_{SF} \cong E_{TRUE} > E_{DEA}$ $E_{SF} > E_{DEA} \cong E_{TRUE}$ $E_{SF} > E_{TRUE} > E_{DEA}$, or $E_{SF} \cong E_{TRUE} > E_{DEA}$, or $E_{TRUE} > E_{SF} > E_{DEA}$, or $E_{SF} > E_{TRUE} > E_{DEA}$
$E_{SF} \cong E_{DEA}$	✓	✗	✗	• all assumptions met	$E_{SF} \cong E_{DEA} \cong E_{TRUE}$
	✗	✓*	✗	• not enough efficient DMUs and FD	$E_{DEA} \cong E_{SF} > E_{TRUE}$
	✗	✗	✓**	• too restrictive RTS in both methods • non-convex PPS • too restrictive RTS in SF and not restrictive enough in DEA or not enough efficient DMUs	$E_{TRUE} > E_{SF} \cong E_{DEA}$ $E_{TRUE} > E_{SF} \cong E_{DEA}$ $E_{DEA} \cong E_{SF} > E_{TRUE}$
$E_{DEA} > E_{SF}$	✓	✗	✗	• incorrect assumption about inefficiency distribution in SF	$\overline{E}_{DEA} \cong \overline{E}_{TRUE} > \overline{E}_{SF}$
	✗	✓*	✗	• too restrictive functional form in SF • not enough efficient DMUs • not enough efficient DMUs and FD in SF	$E_{DEA} \cong E_{TRUE} > E_{SF}$ $E_{DEA} > E_{TRUE} \cong E_{SF}$ $E_{DEA} > E_{SF} > E_{TRUE}$, or $E_{DEA} > E_{TRUE} > E_{SF}$
	✗	✗	✓**	• too restrictive RTS in SF • need a more restrictive RTS assumption in DEA • non-convex PPS • too restrictive RTS in SF and not restrictive enough in DEA • too restrictive RTS in both methods	$E_{DEA} \cong E_{TRUE} > E_{SF}$ $E_{DEA} > E_{TRUE} \cong E_{SF}$ $E_{TRUE} > E_{DEA} > E_{SF}$ $E_{DEA} > E_{SF} > E_{TRUE}$, or $E_{DEA} > E_{TRUE} > E_{SF}$ $E_{TRUE} > E_{DEA} > E_{SF}$

* and **: If the SF method involves functional deviation, there will be some regions where the SF estimates are greater than the true values, and some where they are less than or equal to the true values.

- too restrictive RTS in SF
- non-convex PPS

The questions that need addressing are:

1. Has either of the methods imposed an unnecessary restriction on the returns to scale of the frontier?
2. Is it possible that the PPS could be non-convex?

Firstly, consider the SF method. If the estimating function which was used in the SF method was a very flexible form, e.g. if a translog or Fourier form was used, then it is unlikely that a restrictive assumption about the returns to scale has been imposed in this method. However, if a Cobb-Douglas form was used, then a less restrictive form may lead to better SF estimates.

In the DEA method, if a CRS, NIRS or NDRS function was imposed, then a VRS model could be used to see whether this has any affect on the results.

If neither of the methods has imposed a restrictive assumption about the returns to scale, then the difference in estimates between the two methods could be explained by a non-convex PPS. In this case, the SF method may have allowed for the non-convexity (possible for flexible forms such as the translog), leading to better SF estimates than DEA,

or it may not, in which case either of the methods could be giving closer estimates.

Similarly, the possible differences at small-scale sizes ($E_{DEA} > E_{SF}$) are;

- too restrictive RTS in SF;
- need a more restrictive RTS assumption in DEA (i.e. not enough efficient units for DEA);
- non-convex PPS;
- too restrictive RTS assumption in SF and not restrictive enough in DEA;
- too restrictive RTS in both methods.

As well as addressing the previous two questions, the density of the units in different regions of scale size should be investigated to see whether the DEA method could be enveloping the data too closely. If a NIRS, NDRS or VRS assumption has been used in the DEA method then using a more restrictive model may lead to better results. The tests developed in Chapter 5 allow this assumption of returns to scale to be tested locally.

8.3.2 Example 2

The SF estimates are found to be less than the DEA estimates at extreme input mixes (i.e. a high ratio of input 1 to input 2 or a high ratio

of input 2 to input 1) but greater than the DEA estimates at other mixes which will be called 'central' input mixes.

Look in Table 8-1 again, but this time look at the rows in the table relating to differences across mix. The different possibilities for the extreme input mixes ($E_{DEA} > E_{SF}$) are

- too restrictive functional form in the SF method
- not enough efficient DMUs
- both of the above.

In this case, the questions become

1. Is the functional form in the SF method implicitly imposing any restrictions on the frontier which may explain the differences between the estimates?
2. Is there a region of the frontier where the density of units is very low?

In order to answer the first question, note that the differences between the estimates from the two methods are varying across the input mix. If this is to be explained by variation of fit of the estimating SF function to the true frontier, it must involve an incorrect imposition about the nature of the elasticity of substitution in the SF method. Does the estimating function which is being used, impose a certain elasticity of substitution on the SF method (e.g. a constant elasticity of substitution)? If so, a more flexible form could be used.

If there is a range of input mixes where there are very few DMUs, it is possible that there are not enough efficient units to define the frontier in the DEA method in this range. This will lead to poor efficiency estimates under DEA in this region. Such regions could be identified by eye by plotting the units across the mix in the two-input case. In higher dimensions, more elaborate methods would be needed which have not been developed here.

Now consider the units operating at the central input mixes. The SF estimates are now greater than the DEA estimates. The causes could be;

- not enough efficient units in DEA and functional deviation in SF
- functional deviation in SF (too restrictive functional form).

This means that there must be functional deviation in the SF method. At the central region the only possible true inequalities are $E_{SF} > E_{DEA} > E_{TRUE}$ (if there are very few units in the central input mix region), or $E_{SF} > E_{TRUE} \cong E_{DEA}$.

In both cases the DEA efficiencies are closer to the true efficiencies than the SF efficiencies.

At the extreme input mixes the possibilities are

- $E_{DEA} \cong E_{TRUE} > E_{SF}$ - none of the DEA assumptions are violated;

- $E_{DEA} > E_{SF} > E_{TRUE}$ or $E_{DEA} > E_{TRUE} > E_{SF}$ - there are not enough efficient DMUs for DEA to properly define the frontier at the extreme mixes.

At the extreme input mixes the SF or DEA estimates could be closer to the true efficiencies depending on whether there are enough efficient DMUs at these mixes and whether the restriction which is imposed in the SF method is restricting the estimated elasticity of substitution to be greater than or less than the true elasticity of substitution. The estimates from the SF method should be recalculated using a more flexible functional form.

Note that Table 8-1 does not include more than one assumption being violated in each method, as this would become extremely complicated. However, it is assumed that the effect of one of the assumptions will dominate the others allowing the results in the table to still give some indication of the possible causes even when several assumptions are violated at once.

8.4 Summary

From Table 8-1 it is clear that there are several reasons for differences between the estimates from the methods. However, it is now possible to see what the possible causes of the differences are and, if one of these causes can be identified as the likely reason for the differences, it

is possible to see which of the methods is giving the better estimates in specific regions of the technology.

Suppose for example, that the output efficiencies of a set of 100 schools need to be determined. Once the inputs and outputs have been decided upon, the DEA and SF methods can be applied. The algorithm developed in this chapter can then be used to investigate the performance of the methods. Suppose the inputs are taken to be the total verbal reasoning score on entry, the number of pupils not receiving free school meals and the number of pupils in the school and take the single output to be the total GCSE score for the school. (See Thanassoulis and Dunstan (1994) for a discussion of these variables.)

The five-step procedure given in the introduction to this chapter should be followed:

1. Identify all the assumptions which the methods make.

Firstly, a DEA model must be chosen. A judgement should be made as to the most likely returns to scale of the units. For example, in this case the CRS model seems to be the most likely. A school which has twice as many pupils, with the same percentage receiving free school meals and a verbal reasoning score on entry which is twice as high as a second school should be able to achieve twice the GCSE score of the second school.

An SF model must also be chosen. In this case, as there are three inputs, a Cobb-Douglas function will be much simpler to specify than a translog function and for this reason alone is often chosen in preference. Other assumptions must be made in the SF model; the distributions of the inefficiency and random noise terms. Suppose in this case, the SF model is Cobb-Douglas function with a normal, half-normal error term.

Therefore, the DEA model is assuming no random noise, CRS, a good spread of efficient units across the technology and convexity of the PPS. The SF model is assuming a normally distributed random noise term and a half-normally distributed inefficiency term; a fixed scale elasticity and elasticity of substitution of unity (the nature of the Cobb-Douglas function); no correlation between the inefficiency and the inputs and that the inefficiency is only in the output (total GCSE score).

2. Identify the regions where the methods give different results. Are these regions on input mix or scale size or across the whole technology?

Both models are applied to the data set and the efficiency values for each school are obtained under each method. These efficiency values can then be compared. The ratio of the DEA efficiency estimate to the SF estimate can be plotted across

- scale size in order to identify any assumptions about the nature of returns to scale that do not hold;
- input mix in order to identify any assumptions about the nature of the elasticity of substitution that do not hold;
- and finally, DEA efficiency, to identify any problems with the assumption about the distribution of the inefficiencies in the SF method. The half-normal distribution assumes that more schools will be efficient than inefficient which may not be the case.

3. Try to identify which assumption(s) may not be holding by using the algorithm in Figure 8-1.

If a pattern is observed in any of the cases given in step 2, it is possible to decide which of the assumptions may be leading to these differences as described in Sections 8.3.1 and 8.3.2.

If no pattern can be observed across any of these dimensions, then the general differences can be investigated. If all the DEA efficiency estimates are equal to or slightly greater than the SF estimates then both methods must be performing well. (Recall that some of the DEA estimates will always be slightly greater than the true values due to finite sample error.)

If, on average, the SF estimates are greater than the DEA estimates then it is likely that the data contains random noise.

If neither of these general differences is observed, more than one assumption is being violated in the methods. In this case, it is difficult to identify which of the assumptions do not hold. The éfficiencies could be obtained again from one of the methods using different restrictions to try to eliminate at least one of the violations. For example, the translog function could be used in the SF method to remove some of the restrictive assumptions, and the whole process could be followed again.

4. If possible test the assumptions.

The assumption of CRS in the DEA model can be tested as described in Chapter 5.

5. Decide which of the methods is giving the closer estimates in each region of the technology.

If a pattern can be seen across any of the dimensions in step 2, then Table 8-1 can be used to decide which of the methods is likely to be giving the better estimates in specific ranges of scale size or input mix.

If it is found that the data apparently contains random noise, then it is likely that neither method will be giving good estimates of the true efficiencies. However, it is just as valuable to know when the estimates are unreliable as when they are good. If the schools are assessed against a data set which has high random noise, then some schools

which are given low efficiency estimates will be producing low GCSE grades due to factors which are outside the control of the school. For example, there may be other factors which have not been taken into account which have a large effect on the performance of the pupils, or there may be measurement errors in the data (although, in this case it is difficult to see how large measurement errors could occur). If there is high random noise in the data, it is unfair to use either of the methods to assess the schools.

Once these steps have been followed, the assessor should be able to say confidently that

- the efficiencies of the schools are robust to the estimation technique and there is a high level of confidence about the estimates; or
- for certain schools, which can be identified, the efficiencies cannot be estimated to a high level of accuracy by one or both of the methods; or
- there is a high level of random noise in the data due to factors that have not been considered or measurement errors, and neither method is giving accurate efficiency estimates. In this case the performance of the schools should not be assessed using either of the methods.

Chapter 9

Summary and Conclusions

9.1 Summary

This thesis has compared Data Envelopment Analysis and Stochastic Frontier as methods for measuring relative technical efficiencies. The methods are generally used independently of each other but it has been shown here that by comparing the results from the methods it is possible to gain some insight into the underlying causes for differences between the methods and possibly to be able to say which of the methods is giving better estimates in certain regions of the technology.

Chapter 1 gave an outline of the measurement of technical efficiency and the structure of the two approaches. The different assumptions of the methods were examined in detail in Chapter 2. This chapter also put forward seven hypotheses for how the assumptions of the methods affect the results when they are individually violated. These hypotheses were then illustrated in the next five chapters.

Chapter 3 examined the assumptions about the error terms in each method. DEA does not allow for any random error, which leads to estimates that become increasingly poor as the level of noise increases. This chapter also illustrated that, although the SF method allows for random noise in the data, the method does not split the total error exactly, as the inefficiency term is still conditional on the total error. This leads to the SF method giving estimates that are closer in magnitude to the true values than the DEA estimates for high random

noise. However, the correlation between the SF estimates and the true values is almost as poor for the SF estimates under high random noise as it is for the DEA estimates (see Table 3-3).

Chapters 4, 5 and 6 examined differences between the methods across scale size. In order to do this a definition of scale size in multiple dimensions was required. This was developed in Chapter 4 for the single output case. In Chapter 5, this measure, cross-mix scale size, was used to develop tests for the true nature of the returns to scale of a DEA frontier. Chapter 6 then developed the cross-mix scale size measure to the general case of multiple inputs and outputs. It was shown that measuring the cmss in multiple dimensions required the estimation of an homothetic CRS frontier. In the case of DEA this necessitates a new DEA model which still needs to be developed.

In Chapter 7, differences between the methods were investigated across input mix. The case of two inputs was considered - more research would be needed to generalise to multiple inputs.

Chapter 8 summarised the results from testing the hypotheses outlined in Chapter 2 and developed an algorithm which can be used to identify possible problems in general comparative efficiency assessments.

9.2 Conclusions

This thesis set out to compare DEA and Stochastic Frontiers as alternative methods for estimating the relative efficiencies of a set of decision-making units. Due to the differences between the underlying assumptions of the approaches, the two methods can give very different estimates for some, or all, of the units in the analysis. It has been shown throughout this thesis (see in particular Figures 3-3, 3-4 and 7-3) that neither SF nor DEA universally gives better results than the other method for all data sets. The relative performance of the methods has been shown to be dependent upon the nature of the underlying data set (i.e. the nature of the returns to scale of the production frontier, the elasticity of substitution of the inputs and outputs, the level of random noise in the data, etc.). As it is not possible to investigate these properties of the data set without first imposing a method to estimate the production frontier, it is not possible to tell how a single method is performing by analysing the results from applying it. However, by comparing the results from two different methods, DEA and SF, which have different assumptions about the underlying data, insight can be gained into the nature of the data.

This analysis can help to validate the results from one method. I.e., if the results from one method are compared with the results from the other method and the results are found to be very similar, then it is possible to say that both methods are likely to be giving good estimates

of the true efficiencies. Similarly, if the results from the two methods are compared and the results differ for some units, then these units can be identified as units that possibly have poor efficiency estimates.

If the units which are given very different efficiency estimates under the two methods are in specific regions of the technology, then stronger conclusions can be drawn. It has been shown that the violation of certain assumptions affects the efficiency estimates across specific dimensions of the technology. By using the algorithm given in Chapter 8, it may be possible to identify which of the assumptions of the methods do not hold for a specific data set and perhaps which of the methods is giving the better estimates for units in different regions of the technology.

One direction for future research would be to extend the algorithm to the cases where more than one assumption is violated in each method. This becomes much more complicated as there are so many combinations of possible violations.

The main area for future research is to develop a DEA model which restricts the frontier to be homothetic. This will enable a new cross-mix efficiency to be measured as discussed in Chapter 6.

Appendix 1

Stochastic Frontiers:

technical details

A1.1 Introduction

This appendix gives some of the technical details of the SF method. The next section derives the density function of the composed error term, ε, when the random noise has a normal distribution and the inefficiency term has a half-normal distribution. It is added here for completeness ("the derivation of the density function of ε is straightforward", Aigner, Lovell and Schmidt (1977)). Section A1.3 discusses the Corrected Ordinary Least Squares method and Section A1.4, the Maximum Likelihood method.

A1.2 The density function of ε

The random noise is assumed to be distributed normally, mean zero and standard deviation σ_v. This distribution is given by

$$f(v) = \frac{1}{\sqrt{2\pi\sigma_v^2}} \exp\left[\frac{-v^2}{2\sigma_v^2}\right] \qquad \text{(A1-1)}$$

The inefficiency, in this case, is distributed half-normally. The mean of the underlying normal is zero, and the standard deviation is σ_u. This is given by

$$g(u) = \begin{cases} \frac{2}{\sqrt{2\pi\sigma_u^2}} \exp\left[\frac{-u^2}{2\sigma_u^2}\right] & \text{for } u > 0 \\ 0 & \text{for } u \leq 0 \end{cases} \qquad \text{(A1-2)}$$

As ε is the sum of the random noise and the inefficiency, the distribution of a sum is needed:

$$K_{X+Y}(Z) = \int_{-\infty}^{\infty} K_X(x) K_Y(Z-x) dx \qquad \text{(A1-3)}$$

Define $\sigma^2 = \sigma_u^2 + \sigma_v^2$ and $\lambda = \frac{\sigma_u}{\sigma_v}$

Using this, the probability density function of ε, $h(\varepsilon)$ is given by

$$\begin{aligned} h(\varepsilon) &= \int_0^{\infty} \frac{1}{\pi\sigma_v\sigma_u} \exp\left[\frac{-u^2}{2\sigma_u^2}\right] \exp\left[\frac{-(\varepsilon+u)^2}{2(\sigma^2-\sigma_u^2)}\right] du \\ &= \int_0^{\infty} \frac{1}{\pi\sigma_u\sigma_v} \exp\left[\frac{-\varepsilon^2}{2\sigma^2}\right] \exp\left[-\frac{1}{2}\left\{\frac{u^2}{\sigma_u^2} + \frac{(\varepsilon+u)^2}{\sigma^2-\sigma_u^2} - \frac{\varepsilon^2}{\sigma^2}\right\}\right] du \\ &= \int_0^{\infty} \frac{1}{\pi\sigma_u\sigma_v} \exp\left[\frac{-\varepsilon^2}{2\sigma^2}\right] \exp\left[-\frac{1}{2}\left\{\frac{(u\sigma^2+\varepsilon\sigma_u^2)^2}{\sigma^2\sigma_u^2(\sigma^2-\sigma_u^2)}\right\}\right] du \\ &= \frac{\sqrt{2}}{\sqrt{\pi}\sigma_u\sigma_v} f^*\left(\frac{\varepsilon}{\sigma}\right) \int_0^{\infty} \exp\left[-\frac{1}{2}\left\{\frac{(u\sigma^2+\varepsilon\sigma_u^2)^2}{\sigma^2\sigma_u^2(\sigma^2-\sigma_u^2)}\right\}\right] du \qquad \text{(A1-4)} \end{aligned}$$

Now let $\frac{(u\sigma^2 + \varepsilon\sigma_u^2)^2}{\sigma^2\sigma_u^2(\sigma^2 - \sigma_u^2)} = K^2$

So du = $\frac{\sigma_u\sigma_v}{\sigma}$ dK, and when u=0, K=$\frac{\varepsilon\sigma_u}{\sigma\sigma_v} = \frac{\varepsilon\lambda}{\sigma}$

[Var(K)=1 since K=$\frac{\left\{\frac{(u\sigma^2 + \varepsilon\sigma_u^2)}{\sigma^2}\right\}}{\left\{\frac{\sigma_u\sigma_v}{\sigma}\right\}}$ where $\left\{\frac{\sigma_u^2\sigma_v^2}{\sigma^2}\right\} = \sigma^{*2}$ is the variance of

$\left\{\frac{(u\sigma^2 + \varepsilon\sigma_u^2)}{\sigma^2}\right\}$ = u - μ*.]

Then

$$h(\varepsilon) = \frac{2}{\sigma} f^*\left(\frac{\varepsilon}{\sigma}\right) \int_{\frac{\varepsilon\lambda}{\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}K^2\right) dK$$

$$= \frac{2}{\sigma} f^*\left(\frac{\varepsilon}{\sigma}\right)\left[F^*(K)\right]_{\frac{\varepsilon\lambda}{\sigma}}^{\infty}$$

$$= \frac{2}{\sigma} f^*\left(\frac{\varepsilon}{\sigma}\right)\left(1 - F^*\left(\frac{\varepsilon\lambda}{\sigma}\right)\right)$$

$$= \frac{2}{\sigma} f^*\left(\frac{\varepsilon}{\sigma}\right) F^*\left(\frac{-\varepsilon\lambda}{\sigma}\right) \qquad \text{(A1-5)}$$

A1.2.1 The mean and variance of the distribution of ε

As Aigner, Lovell and Schmidt only quote the results for the mean and variance of the distribution in (A1-5), the proofs can be found below.

The mean of the distribution is given by

$$E(\varepsilon) = \int_0^{\infty} u \frac{\sqrt{2}}{\sqrt{\pi}\sigma_u} \exp\left[\frac{-u^2}{2\sigma_u^2}\right] du$$

$$= \left[-\frac{\sqrt{2}}{\sqrt{\pi}}\sigma_u \exp\left[-\frac{u^2}{2\sigma_u^2}\right]\right]_0^{\infty} \qquad \text{(A1-6)}$$

$$= \frac{\sqrt{2}}{\sqrt{\pi}}\sigma_u$$

The variance can be calculated as follows

$$E(u^2) = \int_0^{\infty} u^2 \frac{\sqrt{2}}{\sqrt{\pi}\sigma_u} \exp\left[\frac{-u^2}{2\sigma_u^2}\right] du$$

Integrate by parts:

$$= \left[-\frac{\sqrt{2}}{\sqrt{\pi}} u\sigma_u \exp\left[-\frac{u^2}{2\sigma_u^2}\right]\right]_0^{\infty} + \int_0^{\infty} \frac{\sqrt{2}}{\sqrt{\pi}}\sigma_u \exp\left[\frac{-u^2}{2\sigma_u^2}\right] du$$

$$= 2\sigma_u^2 [F^*(u)]_0^{\infty} = \sigma_u^2$$

So

$$\text{Var}(u) = \sigma_u^2 - \frac{2}{\pi}\sigma_u^2 = \frac{\pi-2}{\pi}\sigma_u^2$$

So

$$\text{Var}(\varepsilon) = \left(\frac{\pi-2}{\pi}\right)\sigma_u^2 + \sigma_v^2$$

(A1-7)

A1.3 Corrected ordinary least squares

A1.3.1 A deterministic frontier

A production function can be estimated using ordinary least squares (OLS) techniques. However, OLS estimates the mean rather than the maximal output, given inputs. This does not fit with the definition of a production function, so the OLS estimates have to be corrected (Richmond J. (1974)). One way to do this is to estimate by OLS and then to correct the constant term by shifting it up until no residual is positive and one is zero. This method is known as corrected ordinary least squares (COLS) (Olson, Schmidt, and Waldman (1980)).

Consider the model

$$y_i = \alpha + \beta^T x_i - u_i \qquad u_i \geq 0 \text{ for all } i \qquad (A1\text{-}8)$$

This equation may now be estimated by OLS to obtain best linear unbiased and consistent estimates for the β. Note that α is not consistently estimated by OLS.

The model can be rewritten as

$$y_i = (\alpha - \mu) + \Sigma\beta_j x_{ji} - (u_i - \mu) \qquad (A1\text{-}9)$$

where μ is the expected value of the u_i and the new error term has zero mean. The estimates of $(\alpha - \mu)$ and β are now consistent (Richmond (1974)).

A1.3.2 A stochastic frontier

For a COLS estimator for the half-normal stochastic frontier see Olson, Schmidt, and Waldman (1980). The second and third moments of the residuals can be used to calculate σ_u and σ_v in (1-21).

Let $\hat{\mu}_2$ and $\hat{\mu}_3$ be the second and third moments of the OLS residuals. Then the parameters σ_u^2 and σ_v^2 can be estimated using

$$\hat{\sigma}_u^2 = \left[\sqrt{\frac{\pi}{2}}\frac{\pi}{\pi-4}\hat{\mu}_3\right]^{\frac{2}{3}}$$

$$\hat{\sigma}_v^2 = \hat{\mu}_2 - \frac{\pi-2}{\pi}\hat{\sigma}_u^2 \qquad \text{(A1-10)}$$

In this case, the OLS estimates are unbiased and consistent, apart from the constant term which has a bias which is the mean of the composed error - $\sqrt{\frac{2}{\pi}}\sigma_u$ for the normal, half-normal case (see equation (A1-6)).

The COLS residuals can then be calculated by subtracting the estimated bias $\left(\sqrt{\frac{2}{\pi}}\hat{\sigma}_u\right)$ from the OLS constant term.

A1.4 Maximum likelihood estimation

If we make assumptions about u and x and specify a distribution for the u_i in Model 3, then the likelihood function can be derived and the model can be analysed statistically using the maximum likelihood estimators (MLEs) (Jondrow et al. (1982)).

The usual assumptions are that the u_i are independently and identically distributed and that they are independent of the x's.

Many different distributions for the u's can be specified, e.g. half-normal, exponential, gamma. The choice of distribution for u is important because the maximum likelihood estimates depend on it. This is a problem as there do not seem to be good a priori arguments for a particular distribution.

Distributions for the efficiency term

- Aigner, Lovell & Schmidt (1977) consider the half-normal distribution and the exponential distribution for the efficiency term.
- Meeusen & van den Broeck (1977) consider the exponential.

- Most of the literature since has considered the half-normal as it is the easiest to work with computationally - however it does assume that the density is concentrated around zero.
- Stevenson (1980) suggested a general truncated normal distribution allowing a non-zero mode and considered the Gamma distribution.
- Greene (1990) developed a Gamma distribution following from his Gamma frontier model for a deterministic frontier (Greene (1980)). The advantage of this distribution is that its asymmetry is determined by one of its parameters. The disadvantage is the increase in the number of parameters needing to be estimated.

The frontier can be estimated by forming the relevant log-likelihood function and optimising it with respect to the parameters. For example, the loglikelihood function for the normal-half-normal error is given by;

$\ln L(y|\beta,\lambda,\sigma^2) =$

$$N \ln\frac{\sqrt{2}}{\sqrt{\pi}} + N \ln\sigma^{-1} + \sum_{i=1}^{N} \ln[1 - F^*(\varepsilon_i \lambda \sigma^{-1})] - \frac{1}{2\sigma^2}\sum_{i=1}^{N} \varepsilon_i^2 \qquad \text{(A1-11)}$$

where F* is the standard normal distribution function, $\lambda = \sigma_u/\sigma_v$ and $\sigma^2 = \sigma_u^2 + \sigma_v^2$.

Appendix 2

Simulating the Data

A2.1 Introduction

In this appendix will be described how, in general, data is simulated and then the four data generating processes which are used in the main body of the thesis will be described.

In order to test the hypotheses proposed in Chapter 2, it is important that only one assumption in the estimating methods is violated at a time. Therefore, the initial data set should satisfy all the assumptions of both methods. This can be tested by applying both methods to the data and checking that the efficiencies for all units are good estimates of the true efficiencies.

Each assumption must be controlled for. To control for

- **DEA A1. No random noise.**

  Do not add any random noise to the data set.

- **DEA A2. The assumption of CRS or the relaxation of this assumption.**

  Choose the same DEA model as the form of the underlying technology.

- **DEA A3. There is a good spread of efficient units across the whole technology.**

  Use a large enough sample size. (See Banker et al. (1993) for how the sample size affects the performance of the methods.) When testing the hypotheses always use sample sizes that are large

enough for the sample size not to be affecting the results. The input values are generally generated from a uniform distribution in the range [5,15] to ensure that the units are evenly spread across the technology.

Should any of the units be set to be efficient? (Bardhan, Cooper, and Kumbhakar (1994): "...We note that assurances of full efficiency for at least some observations is required to conform to the assumptions usually made in economics. A value of 20% was chosen for mean technical inefficiency since it is consistent with empirical estimates for this parameter as reported in previous DEA studies." In DGPs B and C 20 - 25% of the units are set to be efficient. However, by using an underlying half-normal distribution to generate the data, the majority of the units will be very close to being efficient.

- **DEA A4. Convexity of the production possibility space.**

  Generate the underlying data from a concave function

- **SF A1. Distribution of the inefficiency term.**

  Use the same distribution in the estimation method as used in the data generating process.

Following Banker et al. (1993) data sets here are generated with an average inefficiency of 0.2. This means that when using an underlying half-normal distribution, using the formula u =

$(\sqrt{2}/\sqrt{\pi})^*\sigma$, u = 0.2 (from Appendix 1), the distribution |N(0,0.25)| should be used.

- **SF A2. Form of the production function.**

  Use a flexible enough function to estimate the underlying frontier. In these cases where we know the underlying function, the estimated and true efficiency values can be compared to see whether the function is flexible enough.

- **SF A3. No correlation between the inefficiency and the exogenous variables.**

  Generate the inefficiency independently of the inputs.

- **SF A4. Distribution of the random noise term.**

  Generate the random noise from a normal distribution.

A2.2 Data Generating Process A

This two input, single output set of data is generated with three levels of random noise and two underlying inefficiency distributions. The underlying technology is piecewise log-linear and is defined by:

$$y_{true} = \begin{cases} \gamma_1 x_1^{\alpha_1} x_2^{\beta_1} \text{ for } y_{true} < 14 & \text{region 1} \\ \gamma_2 x_1^{\alpha_2} x_2^{\beta_2} \text{ for } 14 \le y_{true} < 20 & \text{region 2} \\ \gamma_3 x_1^{\alpha_3} x_2^{\beta_3} \text{ for } 20 \le y_{true} < 26 & \text{region 3} \\ \gamma_4 x_1^{\alpha_4} x_2^{\beta_4} \text{ for } 26 \le y_{true} & \text{region 4} \end{cases} \qquad \text{(A2-1)}$$

where y_{true} is the true efficient output, x_1 and x_2 the two inputs and the γ_i defined for continuity.

The α and β are chosen so that the returns to scale vary across the technology.

We have set;

$$\text{returns to scale} = \begin{cases} \alpha_1 + \beta_1 = 2 \text{ for } y_{true} < 14 \\ \alpha_2 + \beta_2 = 1.5 \text{ for } 14 \le y_{true} < 20 \\ \alpha_3 + \beta_3 = 1 \text{ for } 20 \le y_{true} < 26 \\ \alpha_4 + \beta_4 = 0.5 \text{ for } 26 \le y_{true} \end{cases} . \qquad (A2\text{-}2)$$

This gives increasing returns in regions 1 and 2, constant returns in region 3 and decreasing returns in region 4. The returns to scale do not vary across input mix as they do in Banker (1993).

The two inputs are generated from a normal distribution $N(10,2^2)$. This gives the majority of the units in the second and third scale sizes. (Region 1 has 29 units, region 2 has 167, region 3 has 282 and region 4 has 22.) The observed output is generated from the true efficient output by scaling with an efficiency term and a normal random noise term.

Two different inefficiencies are generated: one is half normal, $|N(0,0.2506^2)|$, and the other is uniform, U[0, 0.3].

Two random noise terms are added to the data (see Chapter 3 for an explanation of where these variances come from.)

Low random noise: $N(0, 0.1*\bar{y}/1.96^2)$

High random noise: $N(0, 0.4*\bar{y}/1.96^2)$

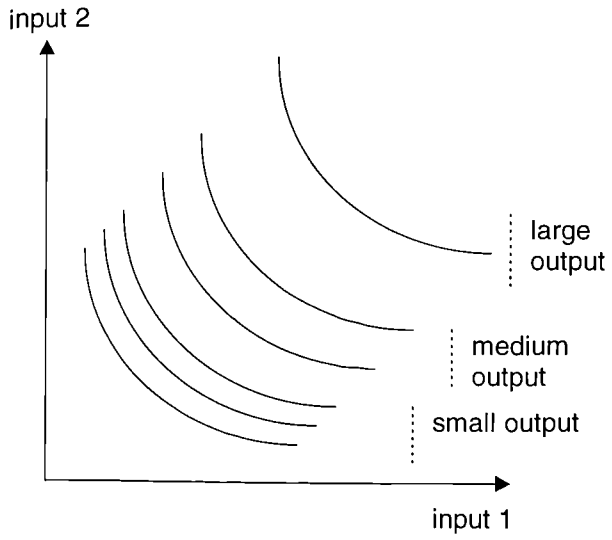
where $\bar{y}$ is the mean of the true efficient outputs.

This method gives us 6 outputs generated from

$$y_{obs} = y_{true}(1 - u_j) + v_i \quad i = 1,2,3,\ j = 1,2 \qquad (A2\text{-}3)$$

where v_1 = zero noise, v_2 = low noise, v_3 = high noise and u_1 = half-normal inefficiency and u_2 = uniform inefficiency.

Figure A2-1. Graph showing how the returns to scale vary across the scale size



To see how this technology looks, consider the graph in Figure A2-1.

The elasticity of substitution is constant across the whole technology - i.e. the 'shape' of the isoquant is constant across the output. The small scale sizes have increasing returns to scale so the isoquants (plotted at unit increase in output) will be close together. As the returns to scale decrease the distance between the isoquants decreases.

A2.3 Data Generating Process B

DGP A had variable returns to scale. In this case the technology has non increasing returns to scale. This property is necessary in Chapter 5 to illustrate that the tests for the nature of the returns to scale can identify 'true' scale inefficiencies and distinguish them from 'apparent' scale inefficiencies.

Once again, the technology for this process is piecewise log-linear. It has CRS for small and medium scale sizes and DRS for large scale sizes. The frontier is given by;

$$y_{true} = \begin{cases} 2X_1^{0.6}X_2^{0.4} \; for \; y_{true} \leq 23.25 \\ \sqrt{46.5}X_1^{0.3}X_2^{0.2} \; for \; y_{true} \geq 23.25 \end{cases} \qquad (A2\text{-}4)$$

This is an homothetic technology by construction.

Each one of the ten data sets constructed from this technology consists of 250 DMUs, their inputs independently generated from U[5,15] and the efficient output from the technology in (A2-4). 20% of the DMUs were generated to operate under DRS and 80% to operate under CRS. The DMUs were then allocated an inefficiency from a half-normal distribution $|N(0,0.25^2)|$ and 20 - 25% of the units were randomly selected to be efficient.

The observed output is then given by

$$y_{obs} = y_{true}\, e^{v-u} \qquad (A2\text{-}5)$$

where e^v is the random noise term, $e^v \sim N(1,0.05^2)$, and e^{-u} is the efficiency term, $1 - e^{-u} \sim |N(0,0.25^2)|$.

A2.4 Data Generating Process C

This data set is generated to illustrate variation of fit across input mix. The function chosen to represent the underlying technology in this case is a Constant Elasticity of Substitution (CES) function, which has the Cobb-Douglas as a special form. This function was chosen because by choosing the parameter values we have easy control over the elasticity of substitution of the function and as it has constant elasticity of substitution we can ensure that the variation of fit is only across the input mix and not across the scale size.

The general form of the CES function[1] under constant returns to scale is given by;

$$y_i^{-\rho} = \gamma^{-\rho}\left(\sum_{k=1}^{n} \delta_k x(k)_i^{-\rho}\right) \qquad \text{(A2-6)}$$

where $x(k)_i$ is the k^{th} input for DMU i, $x(k)_i \geq 0$ and the single output $y_i \geq 0$. We have $\rho \geq 0$, $\gamma > 0$, $\delta_k > 0$, $\forall k$, $\sum\delta_k = 1$, where n is the number of inputs. This function has constant returns to scale. It is possible to specify a more general CES function with increasing or decreasing returns to scale but we are not concerned with that here.

This function has elasticity of substitution σ;

$$\sigma = \frac{\left\{\dfrac{d\left(x_1/x_2\right)}{\left(x_1/x_2\right)}\right\}}{\left\{\dfrac{d\left(dx_1/dx_2\right)}{\left(dx_1/dx_2\right)}\right\}} = \frac{1}{1+\rho}. \qquad \text{(A2-7)}$$

Consider a technology with one output and two inputs: Setting the parameter values to be $\gamma = 1$, $\delta_1 = 0.65$, $\delta_2 = 0.35$, $\rho = 2.33$ gives the underlying technology as:

[1] See (Guilkey, Lovell and Sickles (1983)) for general properties of the CES function.

$$y_{true} = (0.65x_1^{-2.33} + 0.35x_2^{-2.33})^{-(1/2.33)} \qquad (A2\text{-}8)$$

where y_{true} is the quantity of output produced if efficient and x_1 and x_2 are the inputs used.

Our underlying technology in (A2-7) gives us the efficient output levels for given input levels. The *observed* output values are generated by adding a random noise term and an inefficiency term to the efficient output. The observed output, y_{obs}, is generated from;

$$y_{obs} = (1 - u)y_{true} + v \qquad (A2\text{-}9)$$

where u is attributable to inefficiency and v to random noise. This method ensures that the efficiency term takes a value between zero and one which enables the true efficiency terms to be compared easily with the estimated values from SF and DEA.

A2.5 Data Generating Process D

This data set once again has two inputs but this time it also has two outputs. The inputs are generated from U[5,15]. The first output is also generated from a uniform distribution, U[10,30]. The second output is generated from

$$y_2 = \left(\frac{0.5 x_1^{0.6} x_2^{0.4}}{0.2 y_1^{0.5}} \right)^{1/0.6} \qquad \text{(A2-10)}$$

The inefficiency term, u, is distributed as $|N(0,0.25^2)|$. 20% of the units have been set to be efficient. No random noise has been added.

The two observed outputs are then generated from

$$y_{obs} = y_{true} (1 - u) \qquad \text{(A2-11)}$$

where u is the inefficiency term.

Appendix 3

The hypothesis tests

Test 1 (Banker (1993))

Assume that the inefficiencies θ_i come from an exponential distribution for each of the two groups with means $\overline{\theta}_1$ and $\overline{\theta}_2$ respectively. The null hypothesis of no difference between the inefficiencies in the two groups; H_0: $\overline{\theta}_1 = \overline{\theta}_2$ can then be tested using the F distribution with $(2N_1,2N_2)$ degrees of freedom with the test statistic given by $T_{EX} = \frac{\hat{\theta}_1}{\hat{\theta}_2}$ where $\hat{\theta}_i$ is the mean inefficiency in group i.

Test 2 (Banker (1993))

Assume that the inefficiencies θ_i come from a half-normal distribution $|N(0,\sigma_i^2)|$, i = 1,2 for each of the two groups. The null hypothesis of no difference between the inefficiencies in the two groups; H_0: $\overline{\theta}_1 - \overline{\theta}_2 = 0$, can then be tested relative to the F distribution with (N_1,N_2) degrees of freedom with the test statistic given by $T_{HN} = \frac{\sum_{i \in G_1} \theta_i^2 / N_1}{\sum_{i \in G_2} \theta_i^2 / N_2}$.

Test 3 (Banker (1993))

Make no assumption about the inefficiency distribution and use the non-parametric Kolmogorov-Smirnov two-sample test. This tests the

difference between the cumulative distributions of the two groups of inefficiencies. See Siegel and Castellan (1988)[5] for details of the test.

In all of these tests the size of the whole group of DMUs being assessed must be large as the true inefficiency distribution is recovered asymptotically by DEA (see Banker (1993)).

Test 4

Make no assumption about the distribution of the DEA inefficiencies and use the non-parametric Mann Whitney test. We have included this test because it is often used in the literature for testing whether two independent groups come from the same distribution. It is a very powerful test and tests whether one of the groups has inefficiencies which are stochastically larger than those of the other group. Again, see Siegel and Castellan (1988) for details of the test.

Test 5

This is the test for significance of the difference between the sample means in the two groups of inefficiencies. Let the mean inefficiency (that is 1 - DEA efficiency rating) in group i be $\hat{\theta}_i$, i = 1,2. The differences between these mean inefficiencies can then be tested using a test for the difference between the means, the null hypothesis being that there is no difference between the true means; H_0: $\overline{\theta}_1 - \overline{\theta}_2 = 0$. The

test uses the Central Limit Theorem to assume that if G_1 and G_2 are sufficiently large, $\hat{\theta}_i$, i = 1,2 are normally distributed, as is $\hat{\theta}_1 - \hat{\theta}_2$. This is true whatever the distribution of θ_i in G_1 and G_2.

The test statistic is $\frac{\hat{\theta}_1 - \hat{\theta}_2}{SE}$ where SE = $s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ and s_p^2 = $\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$ and s_1, s_2 are the standard deviations of inefficiencies in groups 1 and 2. It is assumed θ_i have the same variance in G_1 and G_2.

Appendix 4

Homotheticity and constant returns to scale

A4.1 Proof of Theorem 1 in Chapter 6

The translog function of multiple inputs and multiple outputs was defined in (6-20) as

$$\ln(r) = \ln(A) + \sum_{i=1}^{m} \alpha_i \ln(x_i) + \sum_{j=1}^{s-1} \beta_j \ln(\theta_j) + \sum_{k=1}^{m}\sum_{l=1}^{m} \gamma_{kl} \ln(x_k)\ln(x_l) + \sum_{p=1}^{s-1}\sum_{q=1}^{m} \delta_{pq} \ln(\theta_p)\ln(x_q) + \sum_{o=1}^{s-1}\sum_{n=1}^{s-1} \lambda_{no} \ln(\theta_o)\ln(\theta_n) \qquad (A4\text{-}1)$$

Now let each of the inputs be scaled by $t \in \Re^+$, then

$$\ln(r(t)) = \ln(A) + \sum_{i=1}^{m} \alpha_i \ln(tx_i) + \sum_{j=1}^{s-1} \beta_j \ln(\theta_j) + \sum_{k=1}^{m}\sum_{l=1}^{m} \gamma_{kl} \ln(tx_k)\ln(tx_l) + \sum_{p=1}^{s-1}\sum_{q=1}^{m} \delta_{pq} \ln(\theta_p)\ln(tx_q) + \sum_{o=1}^{s-1}\sum_{n=1}^{s-1} \lambda_{no} \ln(\theta_o)\ln(\theta_n) \qquad (A4\text{-}2)$$

The elasticity of scale of the function (A4-1) is then given by

$$e(x) = \frac{d\ln(r(t))}{d\ln(t)} \qquad (A4\text{-}3)$$

(see, for example, Varian (1992) page 17).

For constant returns to scale, the elasticity of scale of the function equals unity. So,

$$\frac{d\ln(r(t))}{d\ln(t)} = \sum_{i=1}^{m}\alpha_i + 2\ln(t)\sum_{k=1}^{m}\sum_{l=1}^{m}\gamma_{kl} + \sum_{k=1}^{m}\sum_{l=1}^{m}\gamma_{kl}(\ln(x_k)+\ln(x_l)) + \sum_{p=1}^{s-1}\sum_{q=1}^{m}\delta_{pq}\ln(\theta_p) = 1 \qquad \text{(A4-4).}$$

Now this must hold for *any* values of x_k, x_l, t and θ_p.

Firstly, take the case that t changes to t'. Then (A4-4) becomes

$$\sum_{i=1}^{m}\alpha_i + 2\ln(t')\sum_{k=1}^{m}\sum_{l=1}^{m}\gamma_{kl} + \sum_{k=1}^{m}\sum_{l=1}^{m}\gamma_{kl}(\ln(x_k)+\ln(x_l)) + \sum_{p=1}^{s-1}\sum_{q=1}^{m}\delta_{pq}\ln(\theta_p) = 1 \qquad \text{(A4-5).}$$

Subtracting (A4-4) from (A4-5) gives

$$2(\ln(t') - \ln(t))\sum_{k=1}^{m}\sum_{l=1}^{m}\gamma_{kl} = 0 \qquad \text{(A4-6).}$$

This must hold for any t' so the first restriction must be

- $\sum_{k=1}^{m}\sum_{l=1}^{m}\gamma_{kl} = 0$

Similarly, varying θ_p to θ_p' gives

$$\sum_{i=1}^{m}\alpha_i + 2\ln(t')\sum_{k=1}^{m}\sum_{l=1}^{m}\gamma_{kl} + \sum_{k=1}^{m}\sum_{l=1}^{m}\gamma_{kl}(\ln(x_k)+\ln(x_l)) + \sum_{p=1}^{s-1}\sum_{q=1}^{m}\delta_{pq}\ln(\theta_p') = 1 \qquad (A4\text{-}7).$$

Subtracting (A4-4) from (A4-7) gives

$$\sum_{p=1}^{s-1}\sum_{q=1}^{m}\delta_{pq}(\ln(\theta_p) - \ln(\theta_p')) = 0 \qquad (A4\text{-}8).$$

This must hold for each p so the restriction is

- $\sum_{q=1}^{m}\delta_{pq} = 0$ for p = 1,...,r-1.

Similarly, varying x_k gives the restriction

- $2\gamma_{ii} + \sum_{k\ 1}^{m}\sum_{l=1,l\neq k}^{m}\gamma_{kl} = 0$

And substituting all the restrictions above into (A4-4) gives the final restriction

- $\sum_{i\ 1}^{m}\alpha_i = 1$

References

Afriat, S. N. (1972). "Efficiency Estimation of Production Functions." *International Economic Review* 13, 568-98.

Aigner, D. J., and S. F. Chu. (1968). "On Estimating the Industry Production Function." *American Economic Review* 58, 826-39.

Aigner, D. J., C. A. K. Lovell, and P. Schmidt. (1977). "Formulation and Estimation of Stochastic Frontier Production Function Models." *Journal of Econometrics* 6, 21-37.

Arnold, V. L., I. R. Bardhan, W. W. Cooper, and S. C. Kumbhakar. (1996). "New Uses of DEA and Statistical Regressions for Efficiency Evaluation and Estimation - With an Illustrative Application to Public Secondary Schools in Texas." *Annals of Operations Research* 66, 255-77.

Arrow, K., H. Chenery, B. Minhas, and R. M. Solow. (1961). "Capital-Labor Substitution and Economic Efficiency." *Review of Economics and Statistics* 43, 225-50.

Athanassopoulos, A. D., and E. Thanassoulis. (1995). "Assessing Marginal Impacts of Investments on the Performance of Organisational Units." *International Journal of Production Economics* 39, no. 1/2, 149-64.

Banker, R. D. (1984). "Estimating Most Productive Scale Size Using Data Envelopment Analysis." *European Journal of Operational Research* 17, 35-44.

———. (1993). "Maximum Likelihood, Consistency and Data Envelopment Analysis: A Statistical Foundation." *Management Science* 39, no. 10, 1265-73.

———. (1996). "Hypothesis Tests Using Data Envelopment Analysis." *Journal of Productivity Analysis* 7, 139-59.

Banker, R. D., H. Chang, and W. W. Cooper. (1996). "Simulation Studies of Efficiency, Returns to Scale and Misspecification with Nonlinear Functions in DEA." *Annals of Operations Research* 66, 233-53.

Banker, R. D., A. Charnes, and W. W. Cooper. (1984). "Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis." *Management Science* 30, no. 9, 1078-92.

Banker, R. D., A. Charnes, W. W. Cooper, and A. Maindiratta. (1988). "A Comparison of DEA and Translog Estimates of Production Frontiers Using Simulated Observations from a Known Technology." *Applications of Modern Production Theory: Efficiency and Productivity.* eds. A. Dogramaci, and R. Färe. Boston: Kluwer Academic Publishers.

Banker, R. D., and W. W. Cooper. (1994). "Validation and Generalization of DEA and its Uses." *Top* 2, no. 2, 249-314.

Banker, R. D., V. M. Gadh, and W. L. Gorr. (1993). " A Monte-Carlo Comparison of Two Production Frontier Estimation Methods: Corrected Ordinary Least Squares and Data Envelopment Analysis." *European Journal of Operational Research* 67, no. 3, 332-43.

Banker, R. D., and R. C. Morey. (1986). " The Use of Categorical Variables in Data Envelopment Analysis." *Management Science* 32, no. 12, 1613-27.

Battese, J. E., and G. S. Corra. (1977). "Estimation of a Production Frontier Model: With Application to the Pastoral Zone of Eastern Australia." *Australian Journal of Agricultural Economics* 21, no. 3, 169-79.

Caves, D. W., L. R. Christensen, and W. E. Diewert. (1982). "The Economic Theory of Index Numbers and Measurement of Input, Output, and Productivity." *Econometrica* 50, 1393-414.

Charnes, A., W. W. Cooper, A. Y. Lewin, and L. M. Seiford. (1995). *Data Envelopment Analysis: Theory, Methodology and Applications.* Boston: Kluwer Academic Publishers.

———. (1978). "Measuring the Efficiency of Decision Making Units." *European Journal of Operational Research* 2, no. 6, 429-44.

Charnes, A., W. W. Cooper, and E. Rhodes. (1981). "Evaluating Program and Managerial Efficiency: An Application of Data Envelopment Analysis to Program Follow Through." *Management Science* 27, no. 6, 668-97.

Christiansen, L. R., D. W. Jorgenson, and L. J. Lau. (1971). "Transendental Logarithmic Production Frontiers." *Review of Economics and Statistics* 65, 28-45.

Cobb, C., and P. Douglas. (1928). "A Theory of Production." *American Economic Review* Supplement to Vol.18, 139-65.

Cole, K. C. (1998). *The Universe and The Teacup: The Mathematics of Truth and Beauty*. London: Little, Brown and Company.

Cooper W.W., S. Kumbhakar, R.M. Thrall, and X. Yu (1995). "DEA and Stochastic Frontier Analyses of the 1978 Chinese Economic Reforms." *Socio-Economic Planning Sciences* 29, no. 2, 85-112.

Cooper, W. W., R. G. Thompson, and R. M. Thrall. (1996). "Introduction: Extensions and New Developments in DEA." *Annals of Operations Research* 66, 3-45.

Cooper, W. W., and K. Tone. (1996). "Measures of Inefficiency in Data Envelopment Analysis and Stochastic Frontier Estimation." *European Journal of Operational Research* 99, 72-88.

Debreu, G. (1959). *Theory of Value*. New York: John Wiley & Sons.

Deprins, D., L. Simar, and H. Tulkens. (1984). "Measuring Labor-Efficiency in Post Offices." *The Performance of Public Enterprises: Concepts and measurements*. 243-67. Amsterdam, North-Holland: Elsevier Science Publishers.

Diewert, W. E. (1971). "An Application of the Shephard Duality Theorem: A Generalized Leontief Production Function." *Journal of Political Economy* 79, 481-507.

Färe, R., and D. Primont. (1995). *Multi-Output Production and Duality: Theory and Applications.* Boston: Kluwer Academic Publishers.

Färe, R., and D. Primont. (1995). "On Inverse Homotheticity" *Bulletin of Economic Research* 47, 161-166.

Farrell, M. J. (1957). "The Measurement of Productive Efficiency." *Journal of the Royal Statistical Society, Series A* 120, 253-90.

Forsund, F. R. (1992). "A Comparison of Parametric and Non-parametric Efficiency Measures: The Case of Norwegian Ferries." *Journal of Productivity Analysis* 3, 25-43.

———. (1996). "On the Calculation of the Scale Elasticity in DEA Models." *Journal of Productivity Analysis* 7, 283-302.

Frisch, R. (1965). *Theory of Production.* Dordrecht-Holland.

Gallant, A. R. (1981). "On the Bias in Flexible Functional Forms and an Essentially Unbiased Form." *Journal of Econometrics* 15, 211-45.

———. (1982). "Unbiased Determination of Production Technologies." *Journal of Econometrics* 20, 285-323.

Gong, B-H., and R. C. Sickles. (1992). "Finite Sample Evidence on the Performance of Stochastic Frontiers and Data Envelopment Analysis Using Panel Data." *Journal of Econometrics* 51, 259-84.

Greene, W. H. (1990). "A Gamma-Distributed Stochastic Frontier Model." *Journal of Econometrics* 46, 141-63.

———. *LIMDEP 7 Reference Manual.* Econometric Software, Inc.

———. (1980). "Maximum Likelihood Estimation of Econometric Frontier Functions." *Journal of Econometrics* 13, 27-56.

Guilkey, D. K., C. A. K. Lovell, and R. C. Sickles. (1983). "A Comparison of the Performance of Three Flexible Functional Forms." *International Economic Review* 24, 591-616.

Hanoch, G. (1970). "Hometheticity in Joint Production." *Journal of Economic Theory* 2, 423-26.

Jondrow, J., C. A. K. Lovell, I. S. Materov, and P. Schmidt. (1982). "On the Estimation of Technical Inefficiency in the Stochastic Frontier Production Function Model." *Journal of Econometrics* 19, 233-38.

Jorgenson, D. W., and L. J. Lau. (1974). "The Duality of Technology and Economic Behaviour." *Review of Economic Studies* 61, 181-200.

Kneip, A., B. U. Park, and L. Simar. (forthcoming). "A Note on the Convergence of Nonparametric DEA Efficiency Measures." *Econometric Theory.*

Koopmans, T. C. (1957). *Three Essays on the State of Economic Science.* New York: McGraw Hill.

Lau, L. J. (1972). "Profit Functions of Technologies with Multiple Inputs and Outputs." *Review of Economics and Statistics* 54, 281-89.

Löthgren, M. (1997). "Generalized Stochastic Frontier Production Models." *Economics Letters* 57, 255-59.

Malmquist, S. (1953). "Index Numbers and Indifference Surfaces." *Trabajos De Estatistics* 4, 209-42.

Maniadakis, N., and L. Read. (1997). "A Note on Productivity Measurement with Malmquist Indices." *Warwick Business School Research Paper* No. 278.

McFadden, D. (1978). "Cost, Revenue and Profit Functions." *Production Economics: A Dual Approach to Theory and Applications.* eds M. Fuss, and D. McFadden Amsterdam: North-Holland Publishing Company.

Meeusen, W., and J. van den Broeck. (1977). "Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error." *International Economic Review* 18, 435-44.

Mitchell, K., and N. M. Onvural. (1996). "Economies of Scale and Scope at Large Commercial Banks: Evidence from the Fourier Flexible Functional Form." *Journal of Money, Credit and Banking* 28, no. 2, 178-99.

Olesen, O. B. (1995). "Some Unsolved Problems in Data Envelopment Analysis: A Survey." *International Journal of Production Economics* 39, no. 1/2, 5-36.

Olson, J. A., P. Schmidt, and D. M. Waldman. (1980). "A Monte-Carlo Study of Estimators of Stochastic Frontier Production Functions." *Journal of Econometrics* 13, 67-82.

Petersen, N. C. (1990). "Data Envelopment Analysis on a Relaxed Set of Assumptions." *Management Science* 36, no. 3, 305-14.

Richmond, J. (1974). "Estimating the Efficiency of Production." *International Economic Review* 15, no. 2, 515-21.

Samuelson, P. A. (1966). "The Fundamental Singularity Theorem for Non-Joint Production." *International Economic Review* 7, 34-41.

Schmidt, P. (1985). "Frontier Production Functions." *Econometric Reviews* 4, 289-328.

———. (1976). "On the Statistical Estimation of Parametric Frontier Production Functions" *Review of Economics and Statistics* 58, 238-39.

Seiford, L. M. (1996). "Data Envelopment Analysis: The Evolution of the State-of-the-Art (1978-1995)." *Journal of Productivity Analysis* 7, 99-137.

Seiford, L. M., and R. M. Thrall. (1990). "Recent Developments in DEA: the Mathematical Programming Appoach to Frontier Analysis." *Journal of Econometrics* 46, 7-38.

Sengupta, J. K. (1987). " Data Envelopment Analysis for Efficiency Measurement in the Stochastic Case." *Computers and Operations Research* 14, no. 2, 117-29.

———. (1990). " Structural Efficiency in Stochastic Models of Data Envelopment Analysis." *International Journal of Systems Science* 21, no. 6, 1047-56.

Shephard, R. W. (1970). *Theory of Cost and Production Functions.* Princeton, New Jersey: Princeton University Press.

Simar, L. (1996). "Aspects of Statistical Analysis in DEA-Type Frontier Models." *Journal of Productivity Analysis* 7, 177-85.

Simar, L., and P. Wilson. (1998). "Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models." *Management Science* 44, no. 11, 49-61.

Stevenson, R. E. (1980). "Likelihood Functions for Generalized Stochastic Frontier Estimation." *Journal of Econometrics* 13, 57-66.

Thanassoulis, E., and P. Dunstan (1994). "Guiding Schools to Improved Performance Using Data Envelopment Analysis: An Illustration with Data from a Local Education Authority." *Journal of the Operational Research Society* 45,1247-1262.

Timmer, C. P. (1971). "Using a Probabilistic Frontier Production Function to Measure Technical Efficiency." *Journal of Political Economy* 79, 776-94.

Tulkens, H. (1993). " On FDH Efficiency Analysis: Some Methodological Issues and Applications to Retail Banking, Courts, and Urban Transit." *Journal of Productivity Analysis* 4, 183-210.

Varian, H. R. (1992). *Microeconomic Analysis*. New York: Norton.

Weinstein, M. A. (1964). "The Sum of Values from a Normal and a Truncated Normal Distribution." *Technometrics* 6, 104-5, 469-70.

Zellner, A., J. Kmenta, and J. Dreze. (1966). "Specificaton and Estimation of Cobb-Douglas Production Functions." *Econometrica* 34, 784-95.